

A CONTINUOUS QUADRATIC PROGRAMMING APPROACH TO
TWO-SET GRAPH PARTITIONING

By

SOONCHUL PARK

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

1999

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor, Professor William W. Hager, for his constant guidance and encouragement. Working with him is always a great pleasure. Not only have I benefited from his wide knowledge of mathematics and algorithm programming, but his gentle and sweet personality has affected me in other areas of life as well.

I would like to thank Professor Gang Bao, Yunmei Chen, and Bernard A. Mair for their service on my supervisory committee and their suggestions for improving my mathematical knowledge.

I would also like to thank Professor Panagote M. Pardalos for his service on my supervisory committee and for helping to improve the terminology in this dissertation.

I would also like to thank Professor Timothy A. Davis for his service on my supervisory committee and for providing a computer science perspective in our weekly seminars.

I would like to thank the department of mathematics at the University of Florida for its financial support during the course of my studies.

I am especially grateful to my parents, brother and sisters, and my wife, Jeongmee Lee, and my pretty daughters Iljung and Yeonjung.

TABLE OF CONTENTS

	<u>page</u>
ACKNOWLEDGEMENTS	ii
LIST OF TABLES	v
LIST OF FIGURES	vi
ABSTRACT	vii
CHAPTERS	
1 INTRODUCTION	1
2 GRADIENT PROJECTION METHOD WITH ACTIVE SET STRATEGY	3
1 Optimality Condition	3
2 Procedures to Determine Active Sets	6
2.1 Starting Procedure	6
2.2 Continuing Procedure	9
3 Extreme Point of Convex Set K	18
4 Stopping Criteria for a Given Search Direction	21
3 EXCHANGE THE SUBSETS FROM PARTITIONED GRAPH	23
1 Optimality Condition for Two Partitioned Sets	25
2 Procedures to Determine Active Sets for Two Partitioned Sets	29
2.1 Starting Procedure for Two Partitioned Sets	29
2.2 Continuing Procedure for Two Partitioned Sets	31
4 ALGORITHM	35
1 Initial Point	38
1.1 The Solution of the Ball Centered at \mathbf{x}_c in K	38
1.2 The Solution of the Ball Centered at \mathbf{x}_c Containing K	48
1.3 The Projection on K of the Solution of the Ball Containing K	49
2 Starting Procedure	53
3 Continuing Procedure	54
3.1 Index Decision	54
4 Error Evaluation	55
5 Perturbation Procedure	56
5.1 Perturbation by Using Power Method	57
6 Exchange Method for a Local Minimizer	60

6.1	Kernighan-Lin Method	60
6.2	Improved Kernighan-Lin Method	61
6.3	Formulas for Exchange Method	61
6.4	Exchange Method Procedure	66
7	Stopping Criteria for a Local Minimizer after the Exchange Method .	70
8	Exchange the Subsets from Partitioned Graph	72
5	NUMERICAL RESULT	73
6	CONCLUSION	79
	REFERENCES	80
	BIOGRAPHICAL SKETCH	82

LIST OF TABLES

<u>Table</u>		<u>page</u>
5.1	Comparison of Kernighan's method and the exchange method.	75
5.2	Result of the combination of the gradient projection method, and the exchange method with the initial point $\mathbf{x} = \mathbf{1} \cdot \frac{m}{n}$	76
5.3	LP graph bisection test problems.	77
5.4	G-Set graph bisection test problems.	78

LIST OF FIGURES

<u>Figure</u>	<u>page</u>
2.1 Finite intervals of α	6
2.2 Convergence of $\overline{g_i}$	7
3.1 Partition the subgraphs and generate new partition by exchanging the partitions of the subgraphs	24
4.1 Outline of program	36
4.2 Outline of partition of the original graph	37
4.3 Convergence of λ	51
4.4 Exchange method for a local minimizer.	71
5.1 Example of graph with 20 vertices	74

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

A CONTINUOUS QUADRATIC PROGRAMMING APPROACH TO
TWO-SET GRAPH PARTITIONING

By

Soonchul Park

August 1999

Chairman: William W. Hager
Major Department: Mathematics

We develop a numerical algorithm for solving two-set graph partitioning problems. This algorithm, based on the gradient projection method and analysis of active sets, exploits a continuous quadratic programming formulation of the discrete problem. As a by-product, we obtain an efficient scheme for projecting a point onto the convex set $K = \{\mathbf{0} \leq \mathbf{x} \leq \mathbf{1}, \mathbf{1}^T \mathbf{x} = m\}$ for a given integer m .

CHAPTER 1 INTRODUCTION

Graph partitioning is an important problem that has extensive applications in many areas, including scientific computing, circuit board and micro-chip design, other layout problems (see [15]), and sparse matrix pivoting to get good ordering of the unknowns in order to reduce the number of nonzero entries generated during Gaussian Elimination. The problem is to partition the vertices of a graph into p parts such that the number of edges connecting vertices in different parts is minimized. A good ordering of a sparse matrix dramatically reduces both the amount of memory as well as the time required to solve the system of equations(see [11]). A class of graph partitioning algorithms reduces the size of the graph (*i.e.*, coarsens the graph) by collapsing vertices and edges, partitions the smaller graph, and then uncoarsens it to construct a partition for the original graph. These are called multilevel graph partitioning schemes (see [5], [6]).

The *2-way* graph partitioning problem is defined as following : Given a graph $G = (V, E)$ with $|V| = n$, partition V into two subsets, V_1 and V_2 , such that $V_1 \cap V_2 = \emptyset$ for $|V_1| = m$ and $|V_2| = n - m$ and the number of edges of E whose incident vertices belong to different subsets is minimized. A *2-way* partition of V is commonly represented by a partition vector P of length n , such that for every vertex $v \in V$, $P[v]$ is 0 or 1, indicating the partition set to which vertex v belongs. Given a partition P , the number of edges whose incident vertices belong to different subsets is called the *edge-cut* of the partition.

We use a continuous programming formulation for the following graph partitioning problem : Given a graph with n vertices and a positive integer $m < n$,

partition the vertices into two parts, one with m vertices and the other with $n - m$ vertices, where the partition is chosen to minimize the number of edges whose incident vertices belong to different subsets.

We test our scheme on a number of graphs arising in linear programming and G-sets (<ftp://dollar.biz.uiowa.edu/pub/yyyye/Gset/>). Our experiments show that our scheme produces partitions that are good enough to compete with those produced by the Metis program, and, in some cases, we obtain better partitions.

Let \mathcal{P} denote the set of $n \times n$ permutation matrices. Let G be a graph with n vertices $V = \{1, 2, \dots, n\}$, and let \mathcal{E} denote the set of edges associated with G . The edges are ordered pairs of the form (i, j) where $i, j \in V$. We assume that $(i, i) \notin \mathcal{E}$ and if $(i, j) \in \mathcal{E}$, then $(j, i) \in \mathcal{E}$. The (i, j) element of the adjacency matrix \mathbf{A}_0 is 1 if $(i, j) \in \mathcal{E}$, while it is 0 otherwise. Consider the following two minimization problems:

$$\min_{\mathbf{P} \in \mathcal{P}} \sum_{j=1}^m \sum_{i=m+1}^n (\mathbf{P}^T \mathbf{A}_0 \mathbf{P})_{ij} \quad (1.1)$$

and

$$\begin{aligned} & \min (\mathbf{1} - \mathbf{x})^T \mathbf{A} \mathbf{x} \\ & \text{subject to } \mathbf{0} \leq \mathbf{x} \leq \mathbf{1}, \mathbf{1}^T \mathbf{x} = m \end{aligned} \quad (1.2)$$

where $\mathbf{1}$ is the vector whose entries are all one and $\mathbf{A} = \mathbf{A}_0 + \mathbf{I}$. Let $\mathbf{1}_m$ be the vector whose first m entries are one and whose remaining entries are zero. We know (see [7]) that (1.1) and (1.2) are equivalent in the following sense :

If $\mathbf{x} = \mathbf{P} \mathbf{1}_m$, where $\mathbf{P} \in \mathcal{P}$, is a solution to (1.2), then \mathbf{P} is a solution to (1.1). And conversely, if $\mathbf{P} \in \mathcal{P}$ is a solution to (1.1), then $\mathbf{x} = \mathbf{P} \mathbf{1}_m$ is a solution to (1.2).

CHAPTER 2 GRADIENT PROJECTION METHOD WITH ACTIVE SET STRATEGY

To solve

$$\begin{aligned} & \min (\mathbf{1} - \mathbf{x})^T \mathbf{A} \mathbf{x} \\ & \text{subject to } \mathbf{0} \leq \mathbf{x} \leq \mathbf{1}, \mathbf{1}^T \mathbf{x} = m, \end{aligned}$$

we will use Gradient Projection Method :

$$\begin{aligned} \mathbf{x}(\alpha) &= Proj_K[\mathbf{x}^k - \alpha \nabla f(\mathbf{x}^k)] \\ \mathbf{x}^{k+1} &= \mathbf{x}(\alpha_k) \end{aligned}$$

where $Proj_K$ represents the projection onto a convex set K .

A step-size α_k can be chosen in various ways, for example, Constant step-size, Armijo's rule or Limited minimization method, etc. We will choose α such that

$$f(\mathbf{x}(\alpha_k)) = \min_{\alpha \geq 0} f(\mathbf{x}(\alpha)).$$

We will focus on how to evaluate the projection,

$$\mathbf{x}(\alpha) = Proj_K[\mathbf{x}^k + \alpha \mathbf{g}].$$

where $\mathbf{g} = -\nabla f(\mathbf{x}^k)$. That is,

$$\begin{aligned} & \min \frac{1}{2} \|\mathbf{x} - \mathbf{y}\|^2 \\ & \text{subject to } \mathbf{0} \leq \mathbf{x} \leq \mathbf{1}, \mathbf{1}^T \mathbf{x} = m \end{aligned}$$

where \mathbf{y} is a given vector.

1 Optimality Condition

$$\begin{aligned} & \min \frac{1}{2} \|\mathbf{x} - \mathbf{y}\|^2 \\ & \text{subject to } \mathbf{0} \leq \mathbf{x} \leq \mathbf{1}, \mathbf{1}^T \mathbf{x} = m \end{aligned} \tag{2.1}$$

where $\mathbf{y} \in R^n$ is given. The solution to the problem (2.1) is $Proj_K(\mathbf{y})$. Since $\|\mathbf{x} - \mathbf{y}\|^2$ is a strongly convex function, there exists a unique minimizer and the following first-order optimality conditions hold: There exist $\lambda \in R^1$ and $\mu, \nu \in R^n$ such that

$$\mathbf{x} - \mathbf{y} + \mu - \nu + \lambda \mathbf{1} = \mathbf{0}, \quad (2.2)$$

$$\mathbf{1}^T \mathbf{x} = m, \quad (2.3)$$

$$\mu^T (\mathbf{1} - \mathbf{x}) = 0, \quad \nu^T \mathbf{x} = 0, \quad (2.4)$$

$$\mu \geq \mathbf{0}, \quad \nu \geq \mathbf{0}, \quad (2.5)$$

$$\mathbf{0} \leq \mathbf{x} \leq \mathbf{1}. \quad (2.6)$$

Conversely, if \mathbf{x}, μ, ν and λ satisfy these conditions, then \mathbf{x} is the unique minimizer of (2.1).

For given $U, L \subset \{1, 2, \dots, n\}$, and $F = (U \cup L)^C$, we define

$$x_i = 1 \quad \text{if } i \in U, \quad x_i = 0 \quad \text{if } i \in L \quad (2.7)$$

$$\mu_i = 0, \quad \nu_i = 0 \quad \text{if } i \in F \quad (2.8)$$

$$\lambda = \frac{(|U| - m + \sum_{i \in F} y_i)}{|F|} \quad (2.9)$$

$$x_i = y_i - \lambda \quad \text{if } i \in F \quad (2.10)$$

$$\mu_i = y_i - 1 - \lambda \quad \text{if } i \in U, \quad \nu_i = \lambda - y_i \quad \text{if } i \in L. \quad (2.11)$$

From (2.2), (2.7), and (2.11), we have

$$\mu_i = 0 \text{ for } i \in L \quad \text{and} \quad \nu_i = 0 \text{ for } i \in U.$$

Since if \mathbf{x}, μ, ν and λ satisfies the conditions, (2.2) - (2.6), \mathbf{x} is the unique minimizer of (2.1), if we show that the choices for \mathbf{x}, μ, ν and λ , (2.7) - (2.11) satisfy the conditions (2.2) - (2.6), then \mathbf{x} in (2.7) and (2.10) is the unique minimizer of (2.1). At first, let us show that these choices for \mathbf{x}, μ, ν and λ , satisfy the conditions (2.2), (2.3) and (2.4) :

(2.2). $\mathbf{x} - \mathbf{y} + \mu - \nu + \lambda \mathbf{1} = \mathbf{0}$, i.e. $x_i - y_i + \mu_i - \nu_i + \lambda = 0 \ \forall i$.

For any $i \in U$, $x_i = 1$, $\mu_i = y_i - 1 - \lambda$ and $\nu_i = 0$. Hence

$$x_i - y_i + \mu_i - \nu_i + \lambda = 0.$$

For any $i \in L$, $x_i = 0$, $\nu_i = \lambda - y_i$ and $\mu_i = 0$. Hence

$$x_i - y_i + \mu_i - \nu_i + \lambda = 0.$$

For $i \in F$, $\mu_i = 0$, $\nu_i = 0$ and $x_i = y_i - \lambda$. So,

$$x_i - y_i + \mu_i - \nu_i + \lambda = 0.$$

Therefore,

$$x_i - y_i + \mu_i - \nu_i + \lambda = 0 \ \forall i.$$

(2.3). $\mathbf{1}^T \mathbf{x} = m$, that is, $\sum_{i=1}^n x_i = m$.

$$\begin{aligned} \sum_{i=1}^n x_i &= |U| + \sum_{i \in F} x_i \\ &= |U| + \sum_{i \in F} y_i - |F| \lambda \\ &= |U| + \sum_{i \in F} y_i - |F| \frac{(|U| - m + \sum_{i \in F} y_i)}{|F|} \\ &= m \end{aligned}$$

(2.4). $\mu^T(\mathbf{1} - \mathbf{x}) = 0$ and $\nu^T \mathbf{x} = 0$, that is, $\mu_i(1 - x_i) = 0$ and $\nu_i x_i = 0$ for all i .

$x_i = 1$ for $i \in U$ and $\mu_i = 0$ for $i \notin U$, and $x_i = 0$ for $i \in L$ and $\nu_i = 0$ for $i \notin L$. Hence,

$$\mu_i(1 - x_i) = 0 \quad \text{and} \quad \nu_i x_i = 0 \quad \forall i.$$

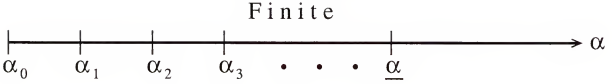
We showed that the choices for \mathbf{x}, μ, ν and λ , (2.7) - (2.11) satisfy the conditions (2.2) - (2.4). Now, if we show that the choices for \mathbf{x}, μ, ν and λ , (2.7) - (2.11) satisfy the conditions (2.5) and (2.6), $\mu \geq \mathbf{0}$ and $\nu \geq \mathbf{0}$, and $\mathbf{0} \leq \mathbf{x} \leq \mathbf{1}$, then the choices for \mathbf{x}, μ, ν and λ , (2.7) - (2.11) satisfy the first-order optimality conditions (2.2) - (2.6),

and \mathbf{x} is the solution of (2.1). Suppose that $\mathbf{y}(\alpha) = \mathbf{x}_0 + \alpha \mathbf{g}$ where $\alpha \geq 0$ scalar, $\mathbf{0} \leq \mathbf{x}_0 \leq \mathbf{1}$ and $\mathbf{1}^T \mathbf{x}_0 = m$. We now show how to compute the associated minimizer $\mathbf{x} = \mathbf{x}(\alpha)$ corresponding to $\mathbf{y} = \mathbf{y}(\alpha)$ in (2.1).

Definition 1 An inequality constraint $g_i(\mathbf{x}) \leq 0$ is said to be *active* at a feasible point \mathbf{x} if $g_i(\mathbf{x}) = 0$ and *inactive* at \mathbf{x} if $g_i(\mathbf{x}) < 0$.

2 Procedures to Determine Active Sets

To determine active sets for $\alpha = 0^+$ and $\alpha = \alpha_j^+$, we need the following two procedures, Starting Procedure and Continuing Procedure.



For each interval, same U and L

Figure 2.1: Finite intervals of α

2.1 Starting Procedure

Let $B_1 = \{i : x_{0i} = 1\}$, $B_0 = \{i : x_{0i} = 0\}$, $\tilde{F}_0 = \{1, 2, \dots, n\}$, and $\tilde{U}_0 = \tilde{L}_0 = \emptyset$. Let

$$\bar{g}_j = \frac{\sum_{i \in \tilde{F}_j} g_i}{|\tilde{F}_j|}, \quad (2.12)$$

$$\gamma_j = \sum_{i \in B_1 \setminus \tilde{U}_j, g_i > \bar{g}_j} g_i - \bar{g}_j \quad (2.13)$$

and

$$\tau_j = \sum_{i \in B_0 \setminus \tilde{L}_j, g_i < \bar{g}_j} \bar{g}_j - g_i. \quad (2.14)$$

Define \tilde{U}_{j+1} and \tilde{L}_{j+1} in the following way. For each j , if $\gamma_j \geq \tau_j$, then put each $i \in B_1 \setminus \tilde{U}_j$ such that $g_i > \bar{g}_j$ into \tilde{U}_{j+1} , and if $\gamma_j < \tau_j$, then put each $i \in B_0 \setminus \tilde{L}_{j+1}$ such that $g_i < \bar{g}_j$ into \tilde{L}_{j+1} . That is,

$$\begin{aligned} \tilde{U}_{j+1} &= \tilde{U}_j \cup \{i \in B_1 \setminus \tilde{U}_j : g_i > \bar{g}_j\} & \text{if } \gamma_j \geq \tau_j \\ \tilde{U}_{j+1} &= \tilde{U}_j & \text{otherwise,} \end{aligned} \quad (2.15)$$

$$\begin{aligned}\tilde{L}_{j+1} &= \tilde{L}_j \cup \{i \in B_0 \setminus \tilde{L}_j : g_i < \overline{g_j}\} & \text{if } \gamma_j < \tau_j \\ \tilde{L}_{j+1} &= \tilde{L}_j & \text{otherwise}\end{aligned}\tag{2.16}$$

and

$$\tilde{F}_{j+1} = \tilde{F}_0 \setminus \left(\tilde{U}_{j+1} \cup \tilde{L}_{j+1} \right).\tag{2.17}$$

Lemma 1 *If $\gamma_j \geq \tau_j$, then $\overline{g_j} \geq \overline{g_k}$ for all $k > j$, and if $\gamma_j < \tau_j$, then $\overline{g_j} \leq \overline{g_k}$ for all $k > j$.*

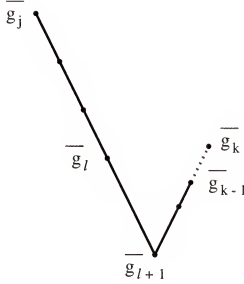


Figure 2.2: Convergence of $\overline{g_i}$

Proof. First, let us consider the case $\gamma_j \geq \tau_j$. Let l be the first index greater than equal to j for which $\gamma_{l+1} < \tau_{l+1}$. Hence, $\gamma_i \geq \tau_i$ for all $j \leq i \leq l$. In (2.15), we remove the indices $t \in B_1 \setminus \tilde{U}_{i-1}$ for which $g_t > \overline{g_i}$ from \tilde{F}_i for each i , $j \leq i \leq l$. Thus, $\overline{g_j} > \cdots > \overline{g_l} > \overline{g_{l+1}}$ and so, if $j < k \leq l+1$ then $\overline{g_j} > \overline{g_k}$. Let k be the first index greater than l for which $\gamma_k \geq \tau_k$. Hence, $\gamma_i < \tau_i$ for all $l+1 \leq i \leq k$. In (2.16), we remove the indices $t \in B_0 \setminus \tilde{L}_{i-1}$ for which $g_t < \overline{g_i}$ from \tilde{F}_i for each i , $l+1 \leq i \leq k$. Thus, $\overline{g_{l+1}} < \cdots < \overline{g_{k-1}} < \overline{g_k}$. We will show that $\overline{g_j} \geq \overline{g_k}$. Since $\overline{g_j} > \cdots > \overline{g_l}$, it is enough to show that $\overline{g_l} \geq \overline{g_k}$ for $k > l+1$. Note that $\overline{g_j} = \overline{g_k}$ holds if $j = l$ and $\gamma_j = \tau_j$ and there is no indices $i \in B_1 \setminus \tilde{U}_{i-1}$ such that $\overline{g_l} > g_i > \overline{g_{l+1}}$.

By way of contradiction, suppose $\overline{g_k} > \overline{g_l}$. Without loss of generality, we can assume $\overline{g_k} > \overline{g_l} \geq \overline{g_{k-1}}$, i.e.

$$\overline{g_{l+1}} < \cdots < \overline{g_{k-1}} \leq \overline{g_l} < \overline{g_k}. \quad (2.18)$$

Let us define the set, N_l as the set of indices in $B_1 \setminus \tilde{U}_l$, such that $g_i \geq \overline{g_l}$, and for $l \leq q \leq k$, M_q as the set of indices in $B_0 \setminus \tilde{L}_q$ such that $g_i < \overline{g_q}$. Then

$$\overline{g_k} = \frac{|\tilde{F}_l| \overline{g_l} - \sum_{N_l} g_i - \sum_{q=l+1}^{k-1} \sum_{i \in M_q} g_i}{|\tilde{F}_k|}. \quad (2.19)$$

Hence,

$$\overline{g_k} - \overline{g_l} = \frac{\sum_{q=l+1}^{k-1} \sum_{i \in M_q} (\overline{g_l} - g_i) - \sum_{N_l} (g_i - \overline{g_l})}{|\tilde{F}_k|}. \quad (2.20)$$

Since $\overline{g_k} > \overline{g_l}$, by (2.20),

$$\sum_{q=l+1}^{k-1} \sum_{i \in M_q} (\overline{g_l} - g_i) - \sum_{N_l} (g_i - \overline{g_l}) > 0. \quad (2.21)$$

On the other hand, since $\gamma_l \geq \tau_l$, we have

$$\sum_{N_l} (g_i - \overline{g_l}) \geq \sum_{M_l} (\overline{g_l} - g_i). \quad (2.22)$$

Hence, from (2.21), we have

$$\sum_{q=l+1}^{k-1} \sum_{i \in M_q} (\overline{g_l} - g_i) - \sum_{M_l} (\overline{g_l} - g_i) > 0. \quad (2.23)$$

Since for $l+1 \leq q \leq k-1$, $B_0 \setminus \tilde{L}_q$ is monotone decreasing and $\overline{g_q} \leq \overline{g_l}$ (2.18), we have

$$M_q \subset M_l$$

and so,

$$\sum_{q=l+1}^{k-1} \sum_{i \in M_q} (\overline{g_l} - g_i) < \sum_{M_l} (\overline{g_l} - g_i). \quad (2.24)$$

This contradicts (2.23). Therefore, $\overline{g_l} \geq \overline{g_k}$ if k satisfies $\gamma_i < \tau_i$ for all $l+1 \leq i \leq k$.

Hence $\overline{g_j} \geq \overline{g_k}$.

Now consider a general $k > j$. We will use mathematical induction to show that $\overline{g_j} \geq \overline{g_k}$ for all $k > j$. Let k_i be the i -th number after j such that $\gamma_{k_i-1} < \tau_{k_i-1}$ and $\gamma_{k_i} \geq \tau_{k_i}$. We have seen that $\overline{g_j} \geq \overline{g_k}$ for $j < k \leq k_1$. Suppose it is true for $j < k \leq k_i$. Then $\overline{g_j} \geq \overline{g_{k_i}}$. Let $k_i < k \leq k_{i+1}$, then by the above argument, we have $\overline{g_{k_i}} \geq \overline{g_k}$. Hence, $\overline{g_j} \geq \overline{g_k}$ for all $k > j$. Similarly, we can show that if $\gamma_j < \tau_j$, then $\overline{g_j} \leq \overline{g_k}$ for all $k > j$. \square

Hence, by Lemma 1,

$$g_i > \overline{g_k} \quad \forall i \in \tilde{U}_j, \quad \forall k \geq j, \quad (2.25)$$

and

$$g_i < \overline{g_k} \quad \forall i \in \tilde{L}_j, \quad \forall k \geq j. \quad (2.26)$$

Let us stop this process if $\gamma_j = \tau_j = 0$. Then since B_0 and B_1 are finite, this must be terminated in finite steps. Let s be the first number such that $\gamma_s = \tau_s = 0$, then define $U_0 = \tilde{U}_s = \tilde{U}_{s+1}$, $L_0 = \tilde{L}_s = \tilde{L}_{s+1}$ and $F_0 = \tilde{F}_s = \tilde{F}_0 \setminus (\tilde{U}_s \cup \tilde{L}_s)$.

Thus, by definitions of \tilde{U}_{j+1} and \tilde{L}_{j+1} , (2.15), (2.16), and (2.25), (2.26),

$$g_k > \frac{\sum_{i \in \tilde{F}_s} g_i}{|\tilde{F}_s|} = \frac{\sum_{i \in F_0} g_i}{|F_0|} \quad \text{if } k \in U_0 = \tilde{U}_s, \quad (2.27)$$

and

$$g_k < \frac{\sum_{i \in \tilde{F}_s} g_i}{|\tilde{F}_s|} = \frac{\sum_{i \in F_0} g_i}{|F_0|} \quad \text{if } k \in L_0 = \tilde{L}_s. \quad (2.28)$$

2.2 Continuing Procedure

$$\begin{aligned} & \min \|\mathbf{x} - \mathbf{y}\|^2 \\ & \text{s.t. } \mathbf{0} \leq \mathbf{x} \leq \mathbf{1}, \quad \mathbf{1}^T \mathbf{x} = m \end{aligned}$$

where $\mathbf{1}^T = (1, 1, \dots, 1)$ and $\mathbf{y} = \mathbf{x}_k + \alpha g$. Let

$$U_j = U(\alpha_j) = \{i : x_i(\alpha_j) = 1\}, \quad L_j = L(\alpha_j) = \{i : x_i(\alpha_j) = 0\}.$$

Lemma 2 *If $U(\alpha_1) = U(\alpha_2)$ and $L(\alpha_1) = L(\alpha_2)$, then for all $\alpha_1 \leq \alpha \leq \alpha_2$,*

$$U(\alpha) = U(\alpha_1) \quad \text{and} \quad L(\alpha) = L(\alpha_1).$$

Proof. Suppose that $U(\alpha_1) = U(\alpha_2)$ and $L(\alpha_1) = L(\alpha_2)$. To show that $U(\alpha) = U(\alpha_1)$ and $L(\alpha) = L(\alpha_1)$ for all $\alpha_1 \leq \alpha \leq \alpha_2$, we need to show that for $0 \leq s \leq 1$,

$$\mathbf{x}(\alpha_1 + s(\alpha_2 - \alpha_1)) = \mathbf{x}(\alpha_1) + s(\mathbf{x}(\alpha_2) - \mathbf{x}(\alpha_1)), \quad (2.29)$$

$$\mu(\alpha_1 + s(\alpha_2 - \alpha_1)) = \mu(\alpha_1) + s(\mu(\alpha_2) - \mu(\alpha_1)), \quad (2.30)$$

$$\nu(\alpha_1 + s(\alpha_2 - \alpha_1)) = \nu(\alpha_1) + s(\nu(\alpha_2) - \nu(\alpha_1)) \quad (2.31)$$

and

$$\lambda(\alpha_1 + s(\alpha_2 - \alpha_1)) = \lambda(\alpha_1) + s(\lambda(\alpha_2) - \lambda(\alpha_1)). \quad (2.32)$$

Suppose that \mathbf{x}, μ, ν and λ satisfy the conditions (2.2) – (2.6), then $\mathbf{x}(\alpha_1 + s(\alpha_2 - \alpha_1))$, $\mu(\alpha_1 + s(\alpha_2 - \alpha_1))$, $\nu(\alpha_1 + s(\alpha_2 - \alpha_1))$, and $\lambda(\alpha_1 + s(\alpha_2 - \alpha_1))$ satisfy the conditions (2.2) – (2.6).

If we show that $\mathbf{x}(\alpha_1) + s(\mathbf{x}(\alpha_2) - \mathbf{x}(\alpha_1))$, $\mu(\alpha_1) + s(\mu(\alpha_2) - \mu(\alpha_1))$, $\nu(\alpha_1) + s(\nu(\alpha_2) - \nu(\alpha_1))$, $\lambda(\alpha_1) + s(\lambda(\alpha_2) - \lambda(\alpha_1))$ satisfy the conditions (2.2) – (2.6), then by the uniqueness of the optimality condition, we have (2.29), (2.30), (2.31), (2.32).

The condition (2.2) is satisfied as follows :

$$\begin{aligned} & x_i(\alpha_1) + s(x_i(\alpha_2) - x_i(\alpha_1)) - (\alpha_1 y_i + s(\alpha_2 y_i - \alpha_1 y_i)) \\ & + (\mu_i(\alpha_1) + s(\mu_i(\alpha_2) - \mu_i(\alpha_1))) + (\nu_i(\alpha_1) + s(\nu_i(\alpha_2) - \nu_i(\alpha_1))) \\ & \quad + \lambda(\alpha_1) + s(\lambda(\alpha_2) - \lambda(\alpha_1)) \\ & = (1 - s)[x_i(\alpha_1) - \alpha_1 y_i + \mu_i(\alpha_1) + \nu_i(\alpha_1) + \lambda(\alpha_1)] \\ & \quad + s[x_i(\alpha_2) - \alpha_2 y_i + \mu_i(\alpha_2) + \nu_i(\alpha_2) + \lambda(\alpha_2)] \\ & = 0 \end{aligned}$$

since $x_i(\alpha_j) - \alpha_j y_i + \mu_i(\alpha_j) + \nu_i(\alpha_j) + \lambda(\alpha_j) = 0$ for all i and $j = 1, 2$.

The condition (2.3) is satisfied as follows :

$$\begin{aligned} \mathbf{1}^T [\mathbf{x}(\alpha_1) + s(\mathbf{x}(\alpha_2) - \mathbf{x}(\alpha_1))] &= \mathbf{1}^T \mathbf{x}(\alpha_1) + s(\mathbf{1}^T \mathbf{x}(\alpha_2) - \mathbf{1}^T \mathbf{x}(\alpha_1)) \\ &= m + s(m - m) \\ &= m. \end{aligned}$$

The condition (2.4) is satisfied as follows :

$$[\mu(\alpha_1) + s(\mu(\alpha_2) - \mu(\alpha_1))]^T [\mathbf{1} - \mathbf{x}(\alpha_1) + s(\mathbf{x}(\alpha_2) - \mathbf{x}(\alpha_1))] = 0$$

$$\Leftrightarrow [\mu_i(\alpha_1) + s(\mu_i(\alpha_2) - \mu_i(\alpha_1))] [1 - x_i(\alpha_1) + s(x_i(\alpha_2) - x_i(\alpha_1))] = 0 \quad \forall i. \quad (2.33)$$

For each i , if $i \in U(\alpha_1) = U(\alpha_2)$, then $x_i(\alpha_1) = x_i(\alpha_2) = 1$, and if $i \notin U(\alpha_1) = U(\alpha_2)$, then $\mu_i(\alpha_1) + s(\mu_i(\alpha_2) - \mu_i(\alpha_1)) = 0$. Hence, (2.33) is satisfied.

$$[\nu(\alpha_1) + s(\nu(\alpha_2) - \nu(\alpha_1))]^T [\mathbf{x}(\alpha_1) + s(\mathbf{x}(\alpha_2) - \mathbf{x}(\alpha_1))] = 0$$

$$\Leftrightarrow [\nu_i(\alpha_1) + s(\nu_i(\alpha_2) - \nu_i(\alpha_1))] [x_i(\alpha_1) + s(x_i(\alpha_2) - x_i(\alpha_1))] = 0 \quad \forall i. \quad (2.34)$$

For each i , if $i \in L(\alpha_1) = L(\alpha_2)$, then $x_i(\alpha_1) = x_i(\alpha_2) = 0$, and if $i \notin L(\alpha_1) = L(\alpha_2)$, then $\nu_i(\alpha_1) + s(\nu_i(\alpha_2) - \nu_i(\alpha_1)) = 0$. Hence, (2.34) is satisfied. Therefore, the condition (2.4) is satisfied.

We now show that the condition (2.5) is satisfied as follows :

$$\mu(\alpha_1) + s(\mu(\alpha_2) - \mu(\alpha_1)) \geq \mathbf{0}$$

$$\Leftrightarrow (1-s)\mu(\alpha_1) + s\mu(\alpha_2) \geq \mathbf{0}. \quad (2.35)$$

Since $\mu(\alpha_1) \geq \mathbf{0}$, $\mu(\alpha_2) \geq \mathbf{0}$ and $0 < s < 1$, (2.35) holds. Also observe that

$$\nu(\alpha_1) + s(\nu(\alpha_2) - \nu(\alpha_1)) \geq \mathbf{0}$$

$$\Leftrightarrow (1-s)\nu(\alpha_1) + s\nu(\alpha_2) \geq \mathbf{0}. \quad (2.36)$$

Since $\nu(\alpha_1) \geq \mathbf{0}$, $\nu(\alpha_2) \geq \mathbf{0}$ and $0 < s < 1$, (2.36) holds. Hence, the condition (2.5) is satisfied.

The condition (2.6) is satisfied as follows :

$$\mathbf{0} \leq \mathbf{x}(\alpha_1) + s(\mathbf{x}(\alpha_2) - \mathbf{x}(\alpha_1)) \leq \mathbf{1}$$

$$\Leftrightarrow 0 \leq x_i(\alpha_1) + s(x_i(\alpha_2) - x_i(\alpha_1)) \leq 1 \quad \forall i$$

$$\Leftrightarrow 0 \leq (1-s)x_i(\alpha_1) + sx_i(\alpha_2) \leq 1 \quad \forall i. \quad (2.37)$$

Since $0 \leq x_i(\alpha_1), x_i(\alpha_2) \leq 1 - s$, $0 \leq (1 - s)x_i(\alpha_1) \leq 1 - s$ and $0 \leq sx_i(\alpha_2) \leq s$.

Hence, (2.37) satisfied and so, (2.6) is satisfied. We showed that

$$\mathbf{x}(\alpha_1) + s(\mathbf{x}(\alpha_2) - \mathbf{x}(\alpha_1)), \mu(\alpha_1) + s(\mu(\alpha_2) - \mu(\alpha_1)), \nu(\alpha_1) + s(\nu(\alpha_2) - \nu(\alpha_1))$$

and

$$\lambda(\alpha_1) + s(\lambda(\alpha_2) - \lambda(\alpha_1)).$$

satisfy the conditions (2.2) – (2.6), and so, by the uniqueness of the optimality condition, (2.29) – (2.32) hold. Thus, for all $\alpha_1 \leq \alpha \leq \alpha_2$,

$$U(\alpha) = U(\alpha_1) \text{ and } L(\alpha) = L(\alpha_1).$$

□

By Lemma 2, we can conclude that there are only finite number of breaks.

Define $\mathcal{U}(\alpha_{j+1})$ and $\mathcal{L}(\alpha_{j+1})$ such that

$$\text{if } \mu_i(\alpha_{j+1}) = 0 \text{ and } x_i(\alpha_{j+1}) = 1, \text{ then } i \in \mathcal{U}(\alpha_{j+1}), \quad (2.38)$$

and

$$\text{if } \nu_i(\alpha_{j+1}) = 0 \text{ and } x_i(\alpha_{j+1}) = 0, \text{ then } i \in \mathcal{L}(\alpha_{j+1}). \quad (2.39)$$

Let

$$F^0 = F_j \cup \mathcal{U}(\alpha_{j+1}) \cup \mathcal{L}(\alpha_{j+1}), \quad U^0 = \emptyset \text{ and } L^0 = \emptyset. \quad (2.40)$$

For $0 \leq l$, define

$$\overline{g}^l = \frac{\sum_{i \in F^l} g_i}{|F^l|}, \quad (2.41)$$

$$\gamma_l = \sum_{i \in \mathcal{U}(\alpha_{j+1}) \setminus U^l, g_i > \overline{g}^l} (g_i - \overline{g}^l) \quad (2.42)$$

and

$$\tau_l = \sum_{i \in \mathcal{L}(\alpha_{j+1}) \setminus L^l, g_i < \overline{g}^l} (\overline{g}^l - g_i). \quad (2.43)$$

Also, define

$$\begin{aligned} U^{l+1} &= U^l \cup \left\{ i \in \mathcal{U}(\alpha_{j+1}) \setminus U^l : g_i > \overline{g^l} \right\} & \text{if } \gamma_l \geq \tau_l \\ U^{l+1} &= U^l & \text{otherwise,} \end{aligned} \quad (2.44)$$

$$\begin{aligned} L^{l+1} &= L^l \cup \left\{ i \in \mathcal{L}(\alpha_{j+1}) \setminus L^l : g_i < \overline{g^l} \right\} & \text{if } \gamma_l < \tau_l \\ L^{l+1} &= L^l & \text{otherwise} \end{aligned} \quad (2.45)$$

and

$$F^{l+1} = F^0 \setminus (U^{l+1} \cup L^{l+1}). \quad (2.46)$$

Let us stop this process if $\gamma_l = \tau_l = 0$. Apply Lemmal with $B_1 = \mathcal{U}(\alpha_{j+1})$, $B_0 = \mathcal{L}(\alpha_{j+1})$, $F_0 = F^0$ and $\gamma_{j+1} = \gamma_l$, $\tau_{j+1} = \tau_l$. Then since $\mathcal{U}(\alpha_{j+1})$ and $\mathcal{L}(\alpha_{j+1})$ are finite, this must be stopped in finite steps. Let s be the first number such that $\gamma_s = \tau_s = 0$, then define U_{j+1} and L_{j+1} with $F_{j+1} = (U_{j+1} \cup L_{j+1})^C$ as follow ;

$$U_{j+1} = (U_j \setminus \mathcal{U}(\alpha_{j+1})) \cup U^s \quad (2.47)$$

and

$$L_{j+1} = (L_j \setminus \mathcal{L}(\alpha_{j+1})) \cup L^s. \quad (2.48)$$

Then as similar to 2.27 and 2.28,

$$g_k > \frac{\sum_{i \in F_{j+1}} g_i}{|F_{j+1}|} \quad \text{if } k \in U^s \quad (2.49)$$

and

$$g_k < \frac{\sum_{i \in F_{j+1}} g_i}{|F_{j+1}|} \quad \text{if } k \in L^s. \quad (2.50)$$

We now define functions $\mathbf{x}(F_j, \alpha_j)$, $\mu(F_j, \alpha_j)$, $\lambda(F_j, \alpha_j)$, with $0 = \alpha_0 < \alpha_1 < \alpha_2 \dots$, where (F_{j+1}, α_{j+1}) are obtained from (F_j, α_j) in the following way ;

$$\lambda(\alpha) = \frac{(|U_j| - m + \sum_{i \in F_j} y_i(\alpha))}{|F_j|}, \quad \text{if } \alpha_j \leq \alpha \leq \alpha_{j+1} \quad (2.51)$$

$$= \frac{|U_j| - m}{|F_j|} + \frac{\sum_{i \in F_j} x_{0i}}{|F_j|} + \frac{\alpha \sum_{i \in F_j} g_i}{|F_j|},$$

$$x_i(\alpha) = y_i(\alpha) - \lambda(F_j, \alpha), \quad \text{if } i \in F_j \text{ and } \alpha_j \leq \alpha \leq \alpha_{j+1} \quad (2.52)$$

$$= x_{0i} + \alpha \left(g_i - \frac{\sum_{i \in F_j} g_i}{|F_j|} \right) - \frac{|U_j| - m}{|F_j|} - \frac{\sum_{i \in F_j} x_{0i}}{|F_j|},$$

$$\begin{aligned}
\mu_i(\alpha) &= y_i(\alpha) - 1 - \lambda(F_j, \alpha), \quad \text{if } i \in U_j \text{ and } \alpha_j \leq \alpha \leq \alpha_{j+1} \\
&= x_{0_i} - 1 + \alpha \left(g_i - \frac{\sum_{i \in F_j} g_i}{|F_j|} \right) - \frac{|U_j| - m}{|F_j|} - \frac{\sum_{i \in F_j} x_{0_i}}{|F_j|},
\end{aligned} \tag{2.53}$$

$$\begin{aligned}
\nu_i(\alpha) &= \lambda(F_j, \alpha) - y_i(\alpha), \quad \text{if } i \in L_j \text{ and } \alpha_j \leq \alpha \leq \alpha_{j+1} \\
&= \frac{|U_j| - m}{|F_j|} + \frac{\sum_{i \in F_j} x_{0_i}}{|F_j|} + \alpha \left(\frac{\sum_{i \in F_j} g_i}{|F_j|} - g_i \right) - x_{0_i},
\end{aligned} \tag{2.54}$$

and

$$\alpha_{j+1} = \min \left\{ \begin{array}{l} \sup\{\alpha : \mu_i(\alpha) > 0 \ \forall i \in U_j, \mu_i(\alpha_j^+) \neq 0\}, \\ \sup\{\alpha : \nu_i(\alpha) > 0 \ \forall i \in L_j, \nu_i(\alpha_j^+) \neq 0\}, \\ \sup\{\alpha : 0 < x_i(F_j, \alpha) < 1 \ \forall i \in F_j, x_i(\alpha_j^+) \neq 0, 1\} \end{array} \right\}. \tag{2.55}$$

By (2.2), (2.52), (2.53), (2.54) and (2.55), we have

$$\mu_i(\alpha) = 0 \quad \text{for } i \notin U_j, \alpha_j \leq \alpha \leq \alpha_{j+1}, \tag{2.56}$$

$$\nu_i(\alpha) = 0 \quad \text{for } i \notin L_j, \alpha_j \leq \alpha \leq \alpha_{j+1}, \tag{2.57}$$

$$x_i(\alpha) = 1 \quad \text{for } i \in U_j, \alpha_j \leq \alpha \leq \alpha_{j+1} \tag{2.58}$$

and

$$x_i(\alpha) = 0 \quad \text{for } i \in L_j, \alpha_j \leq \alpha \leq \alpha_{j+1}. \tag{2.59}$$

Theorem 3 *For all i and α , $x_i(\alpha)$, $\lambda(\alpha)$, $\mu_i(\alpha)$ and $\nu_i(\alpha)$ are continuous piecewise linear functions of α , and $0 \leq x_i(\alpha) \leq 1$.*

Proof. We want to show that $x(F_j, \alpha)$, $\lambda(F_j, \alpha)$ and $\mu(F_j, \alpha)$ are continuous and peicewise linear. Suppose A is the set of indices i such that $x_i = 1$ and D is the set of indices i such that $x_i = 0$ where $A = A_1 \cup A_2$ and $D = D_1 \cup D_2$ with $A_1 = U_j \cap F_{j-1}$, $A_2 = U_{j-1} \cap F_j$, $D_1 = L_j \cap F_{j-1}$, $D_2 = L_{j-1} \cap F_j$. Then

$$|U_j| = |U_{j-1}| + |A_1| - |A_2|, \quad |L_j| = |L_{j-1}| + |D_1| - |D_2|$$

and

$$|F_j| = |F_{j-1}| - (|A_1| - |A_2| + |D_1| - |D_2|).$$

Also, by definition of $x_i(F_j, \alpha)$, (2.52),

$$y_i = 1 + \lambda(F_{j-1}, \alpha_j), \quad \text{if } i \in A$$

and

$$y_i = \lambda(F_{j-1}, \alpha_j), \quad \text{if } i \in D.$$

Hence,

$$\begin{aligned} \lambda(F_{j-1}, \alpha_j) &= \frac{|U_{j-1}| - m + \sum_{i \in F_{j-1}} y_i}{|F_{j-1}|} \\ &= \frac{|U_{j-1}| - m + \sum_{i \in F_j} y_i + \sum_{i \in A_1} y_i - \sum_{i \in A_2} y_i + \sum_{i \in D_1} y_i - \sum_{i \in D_2} y_i}{|F_{j-1}|} \\ &= \frac{|U_j| - m + \sum_{i \in F_j} y_i + \sum_{i \in (A_1 \setminus A_2) \cup (D_1 \setminus D_2)} \lambda(F_{j-1}, \alpha_j)}{|F_{j-1}|}. \end{aligned}$$

Since $(|F_{j-1}| - (|A_1| - |A_2| + |D_1| - |D_2|)) \lambda(F_{j-1}, \alpha_j) = |U_j| - m + \sum_{i \in F_j} y_i \quad \forall j$,

$$\begin{aligned} \lambda(F_{j-1}, \alpha_j) &= \frac{|U_j| - m + \sum_{i \in F_j} y_i}{|F_{j-1}| - (|A_1| - |A_2| + |D_1| - |D_2|)} \\ &= \frac{|U_j| - m + \sum_{i \in F_j} y_i}{|F_j|} \\ &= \lambda(F_j, \alpha_j). \end{aligned}$$

Thus, by definitions of $x_i(\alpha)$ (2.52), $\mu_i(\alpha)$ (2.53) and $\nu_i(\alpha)$ (2.54),

$$x_i(F_{j-1}, \alpha_j) = x_i(F_j, \alpha_j) \quad \forall i \text{ and } j,$$

$$\mu_i(F_{j-1}, \alpha_j) = \mu_i(F_j, \alpha_j) \quad \forall i \text{ and } j,$$

and

$$\nu_i(F_{j-1}, \alpha_j) = \nu_i(F_j, \alpha_j) \quad \forall i \text{ and } j.$$

Hence, $\lambda(\alpha)$, $x(\alpha)\mu(\alpha)$ and $\nu(\alpha)$ are continuous. Moreover, since $x(\alpha)$, $\lambda(\alpha)$, $\mu(\alpha)$ and $\nu(\alpha)$ are piecewise linear, derivatives exist on the interior of any interval (α_j, α_{j+1}) .

Now we want to show that $0 \leq x_i(\alpha) \leq 1$ for all i that is, $\mathbf{0} \leq x(\alpha) \leq \mathbf{1}$ for all $\alpha \geq 0$. If we show that $x_i(0^+) \leq 1$ for $x_{0_i} = 1$, and $x_i(0^+) \geq 0$ for $x_{0_i} = 0$, i.e. for

$k \in F_0 \cap B_1$, $x_k(0^+) \leq 1$ and for $k \in F_0 \cap B_0$, $x_k(0^+) \geq 0$, then by continuity of $\mathbf{x}(\alpha)$, and definition of $\mathbf{x}(\alpha)$ and α_{j+1} , (2.52), (2.55), (2.58), and (2.59), $0 \leq x_i(\alpha) \leq 1$ for all i and $\alpha \geq 0$.

Suppose that $k \in F_0 \cap B_1$. Then since $\gamma_l = 0$, by definition of γ_j (2.13), we have $\{i \in B_1 \setminus \tilde{U}_l : g_i > \bar{g}_l\} = \emptyset$, i.e.,

$$g_k \leq \bar{g}_l = \frac{\sum_{i \in \tilde{F}_l} g_i}{|\tilde{F}_l|} = \frac{\sum_{i \in F_0} g_i}{|F_0|} \quad \forall k \in (B_1 \setminus \tilde{U}_l) = F_0 \cap B_1. \quad (2.60)$$

Hence,

$$x'_k(0^+) = g_k - \frac{\sum_{i \in F_0} g_i}{|F_0|} \leq 0.$$

Since $x_{0_k} = 1$, therefore,

$$x_k(F_0, 0^+) = 1 + 0^+ \left(g_k - \frac{\sum_{i \in F_0} g_i}{|F_0|} \right) \leq 1.$$

Suppose $k \in F_0 \cap B_0$. Then since $\tau_l = 0$, by definition of τ_j (2.14), we have $\{i \in B_0 \setminus \tilde{L}_l : g_i < \bar{g}_l\} = \emptyset$, i.e.

$$g_k \geq \bar{g}_l = \frac{\sum_{i \in \tilde{F}_l} g_i}{|\tilde{F}_l|} = \frac{\sum_{i \in F_0} g_i}{|F_0|} \quad \forall k \in (B_0 \setminus \tilde{L}_l) = F_0 \cap B_0. \quad (2.61)$$

Hence,

$$x'_k(0^+) = g_k - \frac{\sum_{i \in F_0} g_i}{|F_0|} \geq 0.$$

Since $x_{0_k} = 0$, therefore,

$$x_k(F_0, 0^+) = 0 + 0^+ \left(g_k - \frac{\sum_{i \in F_0} g_i}{|F_0|} \right) \geq 0.$$

Hence, by linearity of $\mathbf{x}(\alpha)$, for all $\alpha \geq 0$, $x_k(\alpha) \leq 1$ for $k \in F_0 \cap B_1$ and $x_k(\alpha) \geq 0$ for $k \in F_0 \cap B_0$. Therefore, for all i and $\alpha \geq 0$, $0 \leq x_i(\alpha) \leq 1$. Thus $\mathbf{0} \leq \mathbf{x}(\alpha) \leq \mathbf{1}$ for all $\alpha \geq 0$. \square

So, the condition (2.6) is satisfied. Now if we hold the condition (2.5), then $\mathbf{x}(\alpha)$ is the solution of (2.1). First, let us show that $\mu_i(0) = 0$ and $\nu_i(0) = 0$ for all i . Let $F_l = (B_1 \cup B_0)^C$ where $B_1 = \{i : x_{0_i} = 1\}$, $B_0 = \{i : x_{0_i} = 0\}$.

If $B_1 = \emptyset$, done for B_1 . Suppose $B_1 \neq \emptyset$, then for $k \in B_1$,

$$\mu_k(0) = x_{0_k} - 1 + 0 \cdot \left(g_k - \frac{\sum_{i \in F_I} g_i}{|F_I|} \right) - \frac{|B_1| - m}{|F_I|} - \frac{\sum_{i \in F_I} x_{0_i}}{|F_I|}$$

where $x_{0_k} = 1$, and $-\frac{|B_1| - m}{|F_I|} - \frac{\sum_{i \in F_I} x_{0_i}}{|F_I|} = 0$ since $\sum_{i \in F_I} x_{0_i} + B_1 + B_0 = \sum_{i \in F_I} x_{0_i} + \sum_{i \in B_1} x_{0_i} = \sum_{i=1}^n x_{0_i} = m$. Hence, $\mu_k(0) = 0$ for all $k \in B_1$.

If $B_0 = \emptyset$, done for B_0 . Suppose $B_0 \neq \emptyset$, then for $k \in B_0$,

$$\nu_k(0) = \frac{|B_1| - m}{|F_I|} + \frac{\sum_{i \in F_I} x_{0_i}}{|F_I|} + 0 \cdot \left(\frac{\sum_{i \in F_I} g_i}{|F_I|} - g_k \right) - x_{0_k}$$

where $x_{0_k} = 0$, and $\frac{|B_1| - m}{|F_I|} + \frac{\sum_{i \in F_I} x_{0_i}}{|F_I|} = 0$. Hence $\nu_k(0) = 0$ for all $k \in B_0$.

By definition of μ in F (2.8), $\mu_i(0) = 0$ and $\nu_i(\alpha) = 0$ for $i \in F_I$. Hence, we have

$$\mu_i(0) = 0 \text{ and } \nu_i(0) = 0 \quad \forall i.$$

Now, we want to show that $\mu(\alpha) \geq 0$ and $\nu(\alpha) \geq 0$ for all $\alpha > 0$. We know that $\mu_i(\alpha_j) = 0 \quad \forall i \in F_j = (U_j \cup L_j)^C$. It is sufficient to show that $\mu_i(\alpha) \geq 0 \quad \forall i \in U_j$ and $\nu_i(\alpha) \geq 0$, for all $i \in L_j$, $\alpha \in (\alpha_j, \alpha_{j+1})$ and $j \geq 0$. By definition of α_{j+1} , for all $\alpha \in (\alpha_j, \alpha_{j+1})$,

$$\mu_i(\alpha) > 0 \quad \forall i \in U_j \setminus \mathcal{U}(\alpha_{j+1}) \quad \text{and} \quad \nu_i(\alpha) > 0 \quad \forall i \in L_j \setminus \mathcal{L}(\alpha_{j+1}).$$

Hence by definition of U_{j+1} and L_{j+1} , (2.47) (2.48), it is enough to show that

$$\mu_i(\alpha) > 0 \quad \forall i \in U^s \quad \text{and} \quad \nu_i(\alpha) > 0 \quad \forall i \in L^s \quad \forall \alpha \in (\alpha_j, \alpha_{j+1}).$$

Since μ and ν are linear for the interval $(\alpha_j, \alpha_{j+1}) \forall j \geq 0$, it is sufficient to show that for all j , $\mu'_i(\alpha_j^+) \geq 0 \quad \forall i \in U^s$ and $\nu'_i(\alpha_j^+) \geq 0 \quad \forall i \in L^s$ for all j .

From the definitions of $\mu_i(\alpha)$ and $\nu_i(\alpha)$ (2.53), (2.54), we have

$$\mu'_k(\alpha_j^+) = g_k - \frac{\sum_{i \in F_j} g_i}{|F_j|}$$

and

$$\nu'_k(\alpha_j^+) = \frac{\sum_{i \in F_j} g_i}{|F_j|} - g_k.$$

By (2.27) and (2.28), we have $\mu'_k(0^+) > 0$ and $\nu'_k(0^+) > 0$, and by (2.49) and (2.50), $\mu'_k(\alpha_j^+) > 0$ for all $k \in U^s$ and $\nu'_k(\alpha_j^+) > 0$ for all $k \in L^s$ for all j . Therefore, $\mu_i(\alpha) \geq 0$ and $\nu_i(\alpha) \geq 0$, for all $\alpha \geq 0$ and i . This holds the condition (2.5). Hence $\mathbf{x}(\alpha)$ is the unique minimizer of (2.1).

3 Extreme Point of Convex Set K

Definition 2 A point \mathbf{x} in a convex set C is said to be an *extreme point* of C if there are no two distinct points \mathbf{x}_1 and \mathbf{x}_2 in C such that $\mathbf{x} = \alpha\mathbf{x}_1 + (1 - \alpha)\mathbf{x}_2$ for some α , $0 < \alpha < 1$.

For $K = \{0 \leq \mathbf{x} \leq \mathbf{1}, \mathbf{1}^T \mathbf{x} = m\}$, by the definition of extreme point of a convex set, a point \mathbf{x} such that $x_i = 0$ or 1 for all i is an *extreme point* of K .

Lemma 4 Let \mathbf{b} and \mathbf{c} be the distinct extreme points of K next to an extreme point of K , \mathbf{a} . Then \mathbf{a} , \mathbf{b} and \mathbf{c} are a vector where m components are ones and $n - m$ components, are zeros. Let θ be the angle between $(\mathbf{a} - \mathbf{b})$ and $(\mathbf{a} - \mathbf{c})$. Then $\theta \leq 90^\circ$. And so, for each extreme point of the convex set K , the θ is less than or equal to 90° .

Proof. Let \mathbf{a} be an extreme point of K and \mathbf{b} and \mathbf{c} be the distinct extreme points of K next to \mathbf{a} . Then

$$\cos \theta = \frac{(\mathbf{a} - \mathbf{b})^T (\mathbf{a} - \mathbf{c})}{\|\mathbf{a} - \mathbf{b}\| \|\mathbf{a} - \mathbf{c}\|}. \quad (2.62)$$

Since $\|\mathbf{a} - \mathbf{b}\| \|\mathbf{a} - \mathbf{c}\| > 0$, if $(\mathbf{a} - \mathbf{b})^T (\mathbf{a} - \mathbf{c}) \geq 0$, then $\cos \theta \geq 0$ and so $\theta \leq 90^\circ$.

$$(\mathbf{a} - \mathbf{b})^T (\mathbf{a} - \mathbf{c}) =$$

$$m - |\{j : \mathbf{a}_j = 1, \mathbf{b}_j = 1\}| - |\{j : \mathbf{a}_j = 1, \mathbf{c}_j = 1\}| + |\{j : \mathbf{b}_j = 1, \mathbf{c}_j = 1\}|. \quad (2.63)$$

Since $\{j : \mathbf{a}_j = \mathbf{c}_j = 1\} = \{j : \mathbf{a}_j = \mathbf{c}_j = \mathbf{b}_j = 1\} \cup \{j : \mathbf{a}_j = \mathbf{c}_j = 1, \mathbf{b}_j = 0\}$ and $\{j : \mathbf{b}_j = \mathbf{c}_j = 1\} = \{j : \mathbf{a}_j = 0, \mathbf{b}_j = \mathbf{c}_j = 1\} \cup \{j : \mathbf{a}_j = \mathbf{b}_j = \mathbf{c}_j = 1\}$,

$$\begin{aligned} (2.63) &= m - (|\{j : \mathbf{a}_j = 1, \mathbf{b}_j = 1\}| + |\{j : \mathbf{a}_j = 1, \mathbf{c}_j = 1, \mathbf{b}_j = 0\}|) \\ &\quad + |\{j : \mathbf{a}_j = 0, \mathbf{b}_j = 1, \mathbf{c}_j = 1\}|. \end{aligned} \quad (2.64)$$

Since $\{j : \mathbf{a}_j = 1, \mathbf{b}_j = 0\} = \{j : \mathbf{a}_j = \mathbf{c}_j = 1, \mathbf{b}_j = 0\} \cup \{j : \mathbf{a}_j = 1, \mathbf{c}_j = 0, \mathbf{b}_j = 0\}$,

$$\begin{aligned}
 (2.64) &= m - (|\{j : \mathbf{a}_j = 1, \mathbf{b}_j = 1\}| + |\{j : \mathbf{a}_j = 1, \mathbf{b}_j = 0\}|) \\
 &\quad + |\{j : \mathbf{a}_j = 1, \mathbf{b}_j = 0, \mathbf{c}_j = 0\}| + |\{j : \mathbf{a}_j = 0, \mathbf{b}_j = 1, \mathbf{c}_j = 1\}| \\
 &= m - (|\{j : \mathbf{a}_j = 1, \mathbf{b}_j = 1\}| + |\{j : \mathbf{a}_j = 1, \mathbf{b}_j = 0\}|) \\
 &\quad + |\{j : \mathbf{a}_j \neq \mathbf{b}_j = \mathbf{c}_j\}| \\
 &= m - |\{j : \mathbf{a}_j = 1\}| + |\{j : \mathbf{a}_j \neq \mathbf{b}_j = \mathbf{c}_j\}| \\
 &= |\{j : \mathbf{a}_j \neq \mathbf{b}_j = \mathbf{c}_j\}| \\
 &\geq 0.
 \end{aligned}$$

Hence $\theta \leq 90^\circ$. Since \mathbf{b} and \mathbf{c} were arbitrary for arbitrary fixed \mathbf{a} , we conclude that for each extreme point of K , the θ is less than or equal to 90° . \square

Let $\bar{\mathbf{x}}$ is an extreme point of K , then the normal cone to K at $\bar{\mathbf{x}}$ is

$$N_K(\bar{\mathbf{x}}) = \{\mathbf{p} \in X^* : \langle \mathbf{p}, \mathbf{x} - \bar{\mathbf{x}} \rangle \leq 0 \quad \forall \mathbf{x} \in K\}. \quad (2.65)$$

Lemma 5 *For every extreme point of the convex set K , if the θ is less than or equal to 90° , then $K - \bar{\mathbf{x}} \subset -N_K(\bar{\mathbf{x}})$.*

Proof. Let $\mathbf{w} \in K - \bar{\mathbf{x}}$, then $\mathbf{w} = \mathbf{v} - \bar{\mathbf{x}}$, $\mathbf{v} \in K$. Since the angle of a convex set K is less than or equal to 90° , for all \mathbf{v} and $\mathbf{x} \in K$,

$$\langle \mathbf{v} - \bar{\mathbf{x}}, \mathbf{x} - \bar{\mathbf{x}} \rangle \geq 0.$$

That is,

$$\langle \mathbf{w}, \mathbf{x} - \bar{\mathbf{x}} \rangle \geq 0 \quad \forall \mathbf{w} \in K - \bar{\mathbf{x}}.$$

Hence, $K - \bar{\mathbf{x}} \subset -N_K(\bar{\mathbf{x}})$. \square

Let us define the projection from \mathbb{R}^n on K , as $Proj_K(\mathbf{z}) = \min_{\mathbf{x} \in K} h(\mathbf{x})$ where $h(\mathbf{x}) = \frac{1}{2} \|\mathbf{x} - \mathbf{z}\|^2$. Then $Proj_K(\mathbf{z}) = \mathbf{x}_*$ if and only if $\nabla h(\mathbf{x}_*) \cdot (\mathbf{x} - \mathbf{x}_*) \geq 0 \quad \forall \mathbf{x} \in K$.

That is,

$$\langle \mathbf{x}_* - \mathbf{z}, \mathbf{x} - \mathbf{x}_* \rangle \geq 0 \quad \forall \mathbf{x} \in K. \quad (2.66)$$

Lemma 6 *Let \bar{x} be an extreme point of K . Then $Proj_K(\mathbf{z}) = \bar{x}$ if and only if $\mathbf{z} - \bar{x} \in N_K(\bar{x})$.*

Proof. Let $Proj_K(\mathbf{z}) = \bar{x}$. Then by (2.66),

$$\langle \bar{x} - \mathbf{z}, \mathbf{x} - \bar{x} \rangle \geq 0 \quad \forall \mathbf{x} \in K. \quad (2.67)$$

$$\Rightarrow \langle \mathbf{z} - \bar{x}, \mathbf{x} - \bar{x} \rangle \leq 0 \quad \forall \mathbf{x} \in K. \quad (2.68)$$

Therefore,

$$\mathbf{z} - \bar{x} \in N_K(\bar{x}). \quad (2.69)$$

Conversely, if $\mathbf{z} - \bar{x} \in N_K(\bar{x})$ then by definition of $N_K(\bar{x})$, (2.65), we get the condition (2.66). Hence, $Proj_K(\mathbf{z}) = \bar{x}$. \square

Now, we need to show that if the projection reaches the extreme point with $\underline{\alpha}$, then for any $\alpha \geq \underline{\alpha}$, the projection will be stay at the extreme point of K .

Theorem 7 *If \bar{x} is an extreme point of K and $Proj_K(\mathbf{x}_0 + \underline{\alpha}\mathbf{y}) = \bar{x}$ where $K - \bar{x} \subset -N_K(\bar{x})$ and $\mathbf{x}_0 \in K$, then $Proj_K(\mathbf{x}_0 + \alpha\mathbf{y}) = \bar{x}$ for all $\alpha \geq \underline{\alpha}$.*

Proof. Since $K - \bar{x} \subset -N_K(\bar{x})$,

$$-(\mathbf{x}_0 - \bar{x}) \in N_K(\bar{x}). \quad (2.70)$$

Since $Proj_K(\mathbf{x}_0 + \underline{\alpha}\mathbf{y}) = \bar{x}$, by lemma 6,

$$\mathbf{x}_0 + \underline{\alpha}\mathbf{y} - \bar{x} = \mathbf{z} - \bar{x} \in N_K(\bar{x}). \quad (2.71)$$

where $\mathbf{z} = \mathbf{x}_0 + \underline{\alpha}\mathbf{y}$. We claim that $\mathbf{x}_0 + (\underline{\alpha} + \delta)\mathbf{y} - \bar{x} = \mathbf{z} + \delta\mathbf{y} - \bar{x} \in N_K(\bar{x})$. By (2.70), (2.71) and the definition of $N_K(\bar{x})$, (2.65) we can get

$$\frac{1}{2}(-(\mathbf{x}_0 - \bar{x})) + \frac{1}{2}(\mathbf{z} - \bar{x}) = \frac{1}{2}\underline{\alpha}\mathbf{y} \in N_K(\bar{x}). \quad (2.72)$$

This implies that all multiple of \mathbf{y} is in $N_K(\bar{x})$ i.e. $\delta\mathbf{y} \in N_K(\bar{x})$. Hence, by definition of normal cone to K at \bar{x} , (2.65) and (2.71),

$$\mathbf{x}_0 + (\underline{\alpha} + \delta)\mathbf{y} - \bar{x} \in N_K(\bar{x}).$$

That is,

$$\mathbf{x}_0 + \alpha \mathbf{y} - \bar{\mathbf{x}} \in N_K(\bar{\mathbf{x}}) \quad \forall \alpha \geq \underline{\alpha}.$$

Therefore, by Lemma 6, $Proj_K(\mathbf{x}_0 + \alpha \mathbf{y}) = \bar{\mathbf{x}}$ for all $\alpha \geq \underline{\alpha}$. \square

Hence, we conclude that if the projection on the convex set K reaches an extreme point of K with $\underline{\alpha}$, then the projection on the convex set K will stay there for all $\alpha \geq \underline{\alpha}$. Therefore, if we reach the extreme point of the convex set K , stop marching on α .

4 Stopping Criteria for a Given Search Direction

Definition 3 Let K be the convex set defined by $K = \{\mathbf{0} \leq \mathbf{x} \leq \mathbf{1}, \quad \mathbf{1}^T \mathbf{x} = m\}$. A point $\mathbf{x}(\alpha_j)$ is the *stopping point* of g on K if we have $\alpha_{j+1} = +\infty$.

Theorem 8 Let g be a given direction vector and U_j, L_j be the upper and lower active sets and F_j inactive set at $\alpha = \alpha_j$. Let $F_j \neq \emptyset$. Then $g_i \geq \bar{g} \quad \forall i \in U_j, \quad g_i \leq \bar{g} \quad \forall i \in L_j$ and $g_i = \bar{g} \quad \forall i \in F_j$ where $\bar{g} = \frac{\sum_{F_j} g_i}{|F_j|}$, if and only if $\alpha_{j+1} = +\infty$.

Proof. Since $F_j \neq \emptyset$, there exist $\bar{g} = \frac{\sum_{F_j} g_i}{|F_j|}$.

$$\begin{aligned} g_i = \bar{g} \quad \forall i \in F_j, \alpha \geq \alpha_j &\Leftrightarrow x'_i(\alpha) = g_i - \bar{g} = 0 \quad \forall i \in F_j, \alpha \geq \alpha_j \\ &\Leftrightarrow x_i(\alpha) = x_i(\alpha_j) \quad \forall i \in F_j, \alpha \geq \alpha_j \end{aligned}$$

So, if $0 < x_i(\alpha_j) < 1$, then $0 < x_i(\alpha) < 1$ for all $i \in F_j, \alpha \geq \alpha_j$. Hence,

$$\sup\{\alpha : 0 < x_i(F_j, \alpha) < 1 \quad \forall i \in F_j, x_i(F_j, \alpha_j^+) \neq 0 \text{ or } 1\} = +\infty.$$

$$\begin{aligned} g_i \geq \bar{g} \quad \forall i \in U_j, \alpha \geq \alpha_j &\Leftrightarrow \mu'_i(\alpha) = g_i - \bar{g} \geq 0 \quad \forall i \in U_j, \alpha > \alpha_j \\ &\Leftrightarrow \mu_i(\alpha) \geq \mu_i(\alpha_j) \quad \forall i \in U_j, \alpha \geq \alpha_j \end{aligned}$$

So, if $\mu_i(\alpha_j) > 0$, then $\mu_i(\alpha) > 0$ for all $i \in U_j, \alpha \geq \alpha_j$. Thus,

$$\sup\{\alpha : \mu_i(\alpha) > 0 \quad \forall i \in U_j, \mu_i(F_j, \alpha_j^+) \neq 0\} = +\infty.$$

Similarly,

$$g_i \leq \bar{g} \ \forall i \in L_j \Leftrightarrow \sup\{\alpha : \nu_i(\alpha) > 0 \ \forall i \in L_j, \nu_i(F_j, \alpha_j^+) \neq 0\} = +\infty.$$

Therefore, by the definition of α_{j+1} , $g_i \geq \bar{g} \ \forall i \in U_j$, $g_i \leq \bar{g} \ \forall i \in L_j$ and $g_i = \bar{g} \ \forall i \in F_j$ where $\bar{g} = \frac{\sum_{F_j} g_i}{|F_j|}$, if and only if $\alpha_{j+1} = +\infty$. \square

Lemma 9 F_j can not be an empty set for all j .

Proof. Suppose that j is the first number such that $F_j = \emptyset$. Then $F_{j-1} \neq \emptyset$.

If $x_i = 1$ or $0 \ \forall i \in F_{j-1}$, then stop marching on α and so F_j does not exist. Hence $0 < x_i < 1$ for some i 's in F_{j-1} and $x_i(\alpha_j) = 1$ or $0 \ \forall i \in F_j$.

Recall the definitions in continuing procedure (2.38) - (2.48). By (2.44) and (2.45), either $U^l = U^{l-1}$ or $L^l = L^{l-1}$ for all l . Since $F_j = \emptyset$, there exist s such that $F^s = \emptyset$ and $F^{s-1} \neq \emptyset$. Then either $U^s = U^{s-1}$ or $L^s = L^{s-1}$. Without loss of generality, assume $L^s = L^{s-1}$. Then $U^s \setminus U^{s-1} = F^{s-1}$. Thus, $g_i > \overline{g^{s-1}} \ \forall i \in F^{s-1}$ where $\overline{g^{s-1}} = \frac{\sum_{i \in F^{s-1}} g_i}{|F^{s-1}|}$. It is a contradiction. Therefore $F_j \neq \emptyset$. \square

By this lemma 9, $\bar{g} = \frac{\sum_{F_j} g_i}{|F_j|}$ and $\frac{|U| - m + \sum_{i \in F_j} x_{0i}}{|F_j|}$ are well defined for all j .

Theorem 10 For a given search direction vector g , we must stop marching on α . That is, we reach the stopping point of g on K or the extreme point of the convex set K .

Proof. If we reach stopping point, we stop marching on α . Since $n < \infty$, by Lemma 2, Theorem 8 and lemma 9, we reach the extreme point of K if we do not have the stopping point of g on K . By Theorem 7, we stop marching on α . \square

CHAPTER 3 EXCHANGE THE SUBSETS FROM PARTITIONED GRAPH

After partitioning a graph into two sets, V and W , we can improve by exchanging some components of V and those of W . That is, we can reduce the *edge-cut* between V and W by exchanging some elements of V and those of W . In 1970, Kernighan and Lin [13] introduced this exchange method between two sets. The Kernighan-Lin algorithm uses pair-swapping, exchanging two components, one from each sets.

We develop the generalization of the Kernighan-Lin method from pair-swapping to set-swapping. At first, we change the graph partitioning problem to continuous problem (see [7]). For any m which is less than $\min(|V|, |W|)$, we exchange a set which has m elements of V and a set which has m elements of W . Figure 3.1 gives us the basic idea, and the following theorem is the quadratic problem for the set-swapping method.

Theorem 11

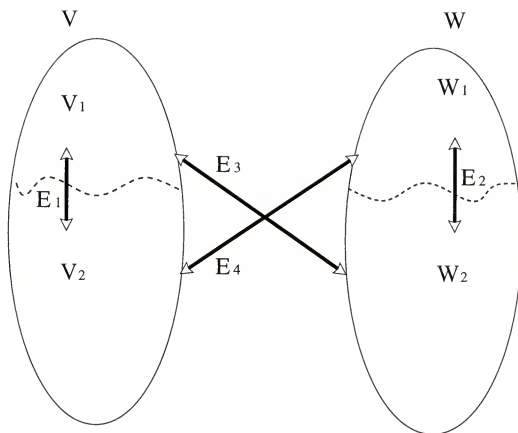
$$\begin{aligned} & \min(\mathbf{1} - \mathbf{x}_1)^T \mathbf{A}_{11} \mathbf{x}_1 + (\mathbf{1} - \mathbf{x}_2)^T \mathbf{A}_{22} \mathbf{x}_2 - (\mathbf{1} - \mathbf{x}_2)^T \mathbf{A}_{21} \mathbf{x}_1 - (\mathbf{1} - \mathbf{x}_1)^T \mathbf{A}_{12} \mathbf{x}_2 \\ & \text{subject to } \mathbf{0} \leq \mathbf{x}_1 \leq \mathbf{1}, \mathbf{0} \leq \mathbf{x}_2 \leq \mathbf{1}, \mathbf{1}^T \mathbf{x}_1 = m = \mathbf{1}^T \mathbf{x}_2, \end{aligned}$$

has a 0/1 solution, where $\mathbf{1}$ is one vector whose size is same as that of \mathbf{x}_i for $i = 1, 2$.

Let V and W be the nodes corresponding to the columns of A_{11} and A_{22} respectively. Let

$$V = V_1 \cup V_2, \quad V_1 \cap V_2 = \emptyset, \quad W = W_1 \cup W_2, \quad W_1 \cap W_2 = \emptyset.$$

where $V_1 = \{i : x_{1i} = 1\}$, $W_1 = \{i : x_{2i} = 1\}$ and $V \cup W = \{1, 2, \dots, n\}$. Then interchanging V_1 and W_1 gives the biggest possible reduction in the edge-cut generated



- E_1 is the number of edge cut between V_1 and V_2
- E_2 is the number of edge cut between W_1 and W_2
- E_3 is the number of edge cut between V_1 and W_2
- E_4 is the number of edge cut between V_2 and W_1

Figure 3.1: Partition the subgraphs and generate new partition by exchanging the partitions of the subgraphs

by swapping sets of size m . Note : The minimum in the Quadratic programming may be greater than or equal to zero in which case no improvement is possible.

Use Gradient Projection Method : Let $\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix}$. Then

$$\mathbf{x}(\alpha) = Proj_{K \times K}[\mathbf{x}^k - \alpha \nabla f(\mathbf{x}^k)],$$

$$\mathbf{x}^{k+1} = \mathbf{x}(\alpha_k)$$

where $Proj_K$ represents the projection onto a convex set K . The projection is

$$\mathbf{x}_1(\alpha) = Proj_K[\mathbf{x}_1^k - \alpha \mathbf{g}_1],$$

$$\mathbf{x}_2(\alpha) = Proj_K[\mathbf{x}_2^k - \alpha \mathbf{g}_2],$$

where $\mathbf{g}_1 = -\nabla_{\mathbf{x}_1} f(\mathbf{x}^k)$ and $\mathbf{g}_2 = -\nabla_{\mathbf{x}_2} f(\mathbf{x}^k)$. That is,

$$\begin{aligned} & \min \frac{1}{2} [\|\mathbf{x}_1 - \mathbf{y}_1\|^2 + \|\mathbf{x}_2 - \mathbf{y}_2\|^2] \\ & \text{subject to } \mathbf{0} \leq \mathbf{x}_1 \leq \mathbf{1}, \mathbf{0} \leq \mathbf{x}_2 \leq \mathbf{1}, \mathbf{1}^T \mathbf{x}_1 = m = \mathbf{1}^T \mathbf{x}_2 \end{aligned}$$

where \mathbf{y}_1 and \mathbf{y}_2 are given vectors.

Let $\mathbf{g} = \begin{pmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \end{pmatrix}$, where

$$\begin{aligned} \mathbf{g}_1 &= -\nabla_{\mathbf{x}_1} f(\mathbf{x}^k) \\ &= -[\mathbf{A}_{11}\mathbf{1} - 2\mathbf{A}_{11}\mathbf{x}_1 - \mathbf{A}_{12}\mathbf{1} + \mathbf{A}_{12}\mathbf{x}_2 + \mathbf{A}_{12}\mathbf{x}_2] \\ &= \mathbf{A}_{11}(2\mathbf{x}_1 - \mathbf{1}) + \mathbf{A}_{12}(\mathbf{1} - 2\mathbf{x}_2) \\ \mathbf{g}_2 &= -\nabla_{\mathbf{x}_2} f(\mathbf{x}^k) \\ &= -[\mathbf{A}_{22}\mathbf{1} - 2\mathbf{A}_{22}\mathbf{x}_2 + \mathbf{A}_{21}\mathbf{x}_1 - \mathbf{A}_{21}\mathbf{1} + \mathbf{A}_{21}\mathbf{x}_1] \\ &= \mathbf{A}_{22}(2\mathbf{x}_2 - \mathbf{1}) + \mathbf{A}_{21}(\mathbf{1} - 2\mathbf{x}_1). \end{aligned}$$

1 Optimality Condition for Two Partitioned Sets

$$\begin{aligned} & \min \frac{1}{2} [\|\mathbf{x}_1 - \mathbf{y}_1\|^2 + \|\mathbf{x}_2 - \mathbf{y}_2\|^2] \\ & \text{subject to } \mathbf{0} \leq \mathbf{x}_1 \leq \mathbf{1}, \mathbf{0} \leq \mathbf{x}_2 \leq \mathbf{1}, \mathbf{1}^T \mathbf{x}_1 = m = \mathbf{1}^T \mathbf{x}_2 \end{aligned} \quad (3.1)$$

where \mathbf{y}_1 and \mathbf{y}_2 are given. That is

$$\begin{aligned} & \min_{\frac{1}{2}\|\mathbf{x} - \mathbf{y}\|^2} \\ & \text{subject to } \mathbf{0} \leq \mathbf{x} \leq \mathbf{1}, \quad \mathbf{1}^T \mathbf{x} = m \end{aligned}$$

where

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix} \in R^n \text{ and } \mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} \in R^n.$$

The solution to the problem (3.1) is $Proj_{K \times K}(\mathbf{y})$, i.e.,

$$\begin{pmatrix} Proj_K(\mathbf{y}_1) \\ Proj_K(\mathbf{y}_2) \end{pmatrix}.$$

Since $\|\mathbf{x}_1 - \mathbf{y}_1\|^2 + \|\mathbf{x}_2 - \mathbf{y}_2\|^2$ is strongly convex function, there exists a unique minimizer and the following first-order optimality conditions hold: There exist $\lambda_1, \lambda_2 \in R^1$ and $\mu_1, \mu_2, \nu_1, \nu_2 \in R^n$ such that

$$\mathbf{x} - \mathbf{y} + \mu - \nu + \lambda \bar{\mathbf{1}} = \mathbf{0},$$

$$\mathbf{1}^T \mathbf{x} = m,$$

$$\mu^T(\mathbf{1} - \mathbf{x}) = 0, \quad \nu^T \mathbf{x} = 0,$$

$$\mu \geq \mathbf{0}, \quad \nu \geq \mathbf{0},$$

$$\mathbf{0} \leq \mathbf{x} \leq \mathbf{1},$$

where $\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix}$, $\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix}$, $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$, $\nu = \begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix}$, $\lambda^T = \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix}$, $\bar{\mathbf{1}} = \begin{bmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} \end{bmatrix}$. That is,

$$\mathbf{x}_t - \mathbf{y}_t + \mu_t - \nu_t + \lambda_t \mathbf{1} = \mathbf{0}, \quad t = 1, 2 \quad (3.2)$$

$$\mathbf{1}^T \mathbf{x}_t = m, \quad t = 1, 2 \quad (3.3)$$

$$\mu_t^T(\mathbf{1} - \mathbf{x}_t) = 0, \quad \nu_t^T \mathbf{x}_t = 0, \quad t = 1, 2 \quad (3.4)$$

$$\mu_t \geq \mathbf{0}, \quad \nu_t \geq \mathbf{0}, \quad t = 1, 2 \quad (3.5)$$

$$\mathbf{0} \leq \mathbf{x}_t \leq \mathbf{1}, \quad t = 1, 2 \quad (3.6)$$

Conversely, if $\mathbf{x}_t, \mu_t, \nu_t$ and λ_t for $t = 1, 2$ satisfies these conditions, then $\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix}$ is the unique minimizer of (3.1).

For given $U_1, L_1 \subset V, F_1 = V \setminus (U_1 \cup L_1)$, and $U_2, L_2 \subset W, F_2 = W \setminus (U_2 \cup L_2)$, we define

$$x_{t_i} = 1 \text{ if } i \in U_t, \quad x_{t_i} = 0 \text{ if } i \in L_t \text{ for } t = 1, 2, \quad (3.7)$$

$$\mu_{t_i} = 0, \quad \nu_{t_i} = 0 \text{ if } i \in F_t \text{ for } t = 1, 2, \quad (3.8)$$

$$\lambda_t = \frac{(|U_t| - m + \sum_{i \in F_t} y_{t_i})}{|F_t|} \text{ for } t = 1, 2, \quad (3.9)$$

$$x_{t_i} = y_{t_i} - \lambda_t \text{ if } i \in F_t \text{ for } t = 1, 2, \quad (3.10)$$

and

$$\mu_{t_i} = y_{t_i} - 1 - \lambda_t \text{ if } i \in U_t, \quad \nu_{t_i} = \lambda_t - y_{t_i} \text{ if } i \in L_t \text{ for } t = 1, 2 \quad (3.11)$$

From (3.2), (3.7) and (3.11), we have, for $t = 1, 2$,

$$\mu_{t_i} = 0 \text{ if } i \in L_t \text{ and } \nu_{t_i} = 0 \text{ if } i \in U_t.$$

We now show that these choices for $\mathbf{x}_1, \mathbf{x}_2, \mu_1, \mu_2, \nu_1, \nu_2, \lambda_1$ and λ_2 satisfy the conditions (3.2), (3.3) and (3.4) as following ;

$$(3.2): \quad \begin{aligned} x_{1_i} - y_{1_i} + \mu_{1_i} - \nu_{1_i} + \lambda_1 &= 0 \quad \forall i \in V, \\ x_{2_i} - y_{2_i} + \mu_{2_i} - \nu_{2_i} + \lambda_2 &= 0 \quad \forall i \in W. \end{aligned}$$

For $i \in U_t$, since $x_{t_i} = 1, \mu_{t_i} = y_{t_i} - 1 - \lambda_t$ and $\nu_{t_i} = 0$ for $t = 1, 2$,

$$x_{t_i} - y_{t_i} + \mu_{t_i} - \nu_{t_i} + \lambda_t = 0 \text{ for } t = 1, 2.$$

For $i \in L_t$, $x_{t_i} = 0$, $\nu_{t_i} = \lambda_t - y_{t_i}$ and $\mu_{t_i} = 0$ for $t = 1, 2$. So,

$$x_{t_i} - y_{t_i} + \mu_{t_i} - \nu_{t_i} + \lambda_t = 0 \text{ for } t = 1, 2.$$

For $i \in F_t$, $\mu_{t_i} = 0, \nu_{t_i} = 0$ and $x_{t_i} = y_{t_i} - \lambda_t$ for $t = 1, 2$. Hence,

$$x_{t_i} - y_{t_i} + \mu_{t_i} - \nu_{t_i} + \lambda_t = 0 \text{ for } t = 1, 2.$$

Therefore,

$$x_{1_i} - y_{1_i} + \mu_{1_i} - \nu_{1_i} + \lambda_1 = 0 \quad \forall i \in V$$

and

$$x_{2_i} - y_{2_i} + \mu_{2_i} - \nu_{2_i} + \lambda_2 = 0 \quad \forall i \in W.$$

$$(3.3) : \sum_{i=1}^{|V|} x_{1_i} = m \text{ and } \sum_{i=1}^{|W|} x_{2_i} = m.$$

$$\begin{aligned} \sum_{i=1}^{|V|} x_{1_i} &= |U_1| + \sum_{i \in F_1} x_{1_i} \\ &= |U_1| + \sum_{i \in F_1} y_{1_i} - |F_1| \lambda_1 \\ &= |U_1| + \sum_{i \in F_1} y_{1_i} - |F_1| \frac{(|U_1| - m + \sum_{i \in F_1} y_{1_i})}{|F_1|} \\ &= m. \end{aligned}$$

$$\begin{aligned} \sum_{i=1}^{|W|} x_{2_i} &= |U_2| + \sum_{i \in F_2} x_{2_i} \\ &= |U_2| + \sum_{i \in F_2} y_{2_i} - |F_2| \lambda_2 \\ &= |U_2| + \sum_{i \in F_2} y_{2_i} - |F_2| \frac{(|U_2| - m + \sum_{i \in F_2} y_{2_i})}{|F_2|} \\ &= m. \end{aligned}$$

$$(3.4) : \mu_{1_i}(1 - x_{1_i}) = 0, \nu_{1_i}x_{1_i} = 0 \text{ for all } i \in V, \text{ and } \mu_{2_i}(1 - x_{2_i}) = 0, \nu_{2_i}x_{2_i} = 0 \text{ for all } i \in W.$$

Since $x_{t_i} = 1$ for $i \in U_t$, $\mu_{t_i} = 0$ for $i \notin U_t$, $x_{t_i} = 0$ for $i \in L_t$ and $\nu_{t_i} = 0$ for $i \notin L_t$ for $t = 1, 2$,

$$\mu_{1_i}(1 - x_{1_i}) = 0 \text{ and } \nu_{1_i}x_{1_i} = 0 \quad \forall i \in V$$

and

$$\mu_{2_i}(1 - x_{2_i}) = 0 \text{ and } \nu_{2_i}x_{2_i} = 0 \quad \forall i \in W.$$

Now, if we show that the conditions (3.5) and (3.6), $\mu_1, \mu_2 \geq 0$ and $\nu_1, \nu_2 \geq 0$, and $0 \leq x_1, x_2 \leq 1$, are satisfied, then the choices for x, μ, ν and λ , (3.7) – (3.10), satisfy

the first-order optimality conditions (3.2) – (3.6), and $\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix}$ is the solution of the quadratic problem (3.1). Suppose that $\mathbf{y}(\alpha) = \mathbf{x}^0 + \alpha \mathbf{g}$, that is, $\mathbf{y}_1(\alpha) = \mathbf{x}_1^0 + \alpha \mathbf{g}_1$ and $\mathbf{y}_2(\alpha) = \mathbf{x}_2^0 + \alpha \mathbf{g}_2$ where $\alpha \geq 0$ scalar, $\mathbf{0} \leq \mathbf{x}_1^0, \mathbf{x}_2^0 \leq \mathbf{1}$ and $\mathbf{1}^T \mathbf{x}_1^0 = m = \mathbf{1}^T \mathbf{x}_2^0$.

2 Procedures to Determine Active Sets for Two Partitioned Sets

To determine active sets for $\alpha = 0^+$ and $\alpha = \alpha_j^+$, we need the following two procedures, Starting Procedure and Continuing Procedure.

2.1 Starting Procedure for Two Partitioned Sets

Let $B_1^1 = \{i : x_{1i}^0 = 1\}$, $B_1^0 = \{i : x_{1i}^0 = 0\}$, $B_2^1 = \{i : x_{2i}^0 = 1\}$, $B_2^0 = \{i : x_{2i}^0 = 0\}$ and $\tilde{F}_{10} = V$, $\tilde{F}_{20} = W$ i.e., $\tilde{U}_{10} = \tilde{L}_{10} = \emptyset$ and $\tilde{U}_{20} = \tilde{L}_{20} = \emptyset$. Let

$$\overline{g_{1j}} = \frac{\sum_{i \in \tilde{F}_{1j}} g_{1i}}{|\tilde{F}_{1j}|}, \quad \overline{g_{2j}} = \frac{\sum_{i \in \tilde{F}_{2j}} g_{2i}}{|\tilde{F}_{2j}|}, \quad (3.12)$$

$$\gamma_{1j} = \sum_{\{i \in B_1^1 \setminus \tilde{U}_{1j}, g_{1i} > \overline{g_{1j}}\}} g_{1i} - \overline{g_{1j}}, \quad \gamma_{2j} = \sum_{\{i \in B_2^1 \setminus \tilde{U}_{2j}, g_{2i} > \overline{g_{2j}}\}} g_{2i} - \overline{g_{2j}}, \quad (3.13)$$

and

$$\tau_{1j} = \sum_{\{i \in B_1^0 \setminus \tilde{L}_{1j}, g_{1i} < \overline{g_{1j}}\}} \overline{g_{1j}} - g_{1i}, \quad \tau_{2j} = \sum_{\{i \in B_2^0 \setminus \tilde{L}_{2j}, g_{2i} < \overline{g_{2j}}\}} \overline{g_{2j}} - g_{2i}. \quad (3.14)$$

Define $\tilde{U}_{1,j+1}$, $\tilde{L}_{1,j+1}$, $\tilde{U}_{2,j+1}$ and $\tilde{L}_{2,j+1}$. For each $j \in N$, if $\gamma_{1j} \geq \tau_{1j}$, then put each $i \in B_1^1 \setminus \tilde{U}_{1j}$ such that $g_{1i} > \overline{g_{1j}}$ into $\tilde{U}_{1,j+1}$, and if $\gamma_{1j} < \tau_{1j}$, then put each $i \in B_1^0 \setminus \tilde{L}_{1,j+1}$ such that $g_{1i} < \overline{g_{1j}}$ into $\tilde{L}_{1,j+1}$. For each $j \in N$, if $\gamma_{2j} \geq \tau_{2j}$, then put each $i \in B_2^1 \setminus \tilde{U}_{2j}$ such that $g_{2i} > \overline{g_{2j}}$ into $\tilde{U}_{2,j+1}$, and if $\gamma_{2j} < \tau_{2j}$, then put each $i \in B_2^0 \setminus \tilde{L}_{2,j+1}$ such that $g_{2i} < \overline{g_{2j}}$ into $\tilde{L}_{2,j+1}$. That is,

$$\begin{aligned} \tilde{U}_{1,j+1} &= \tilde{U}_{1j} \cup \{i \in B_1^1 \setminus \tilde{U}_{1j} : g_{1i} > \overline{g_{1j}}\} & \text{if } \gamma_{1j} \geq \tau_{1j}, \\ \tilde{U}_{1,j+1} &= \tilde{U}_{1j} & \text{otherwise,} \\ \tilde{U}_{2,j+1} &= \tilde{U}_{2j} \cup \{i \in B_2^1 \setminus \tilde{U}_{2j} : g_{2i} > \overline{g_{2j}}\} & \text{if } \gamma_{2j} \geq \tau_{2j}, \\ \tilde{U}_{2,j+1} &= \tilde{U}_{2j} & \text{otherwise.} \end{aligned} \quad (3.15)$$

$$\begin{aligned} \tilde{L}_{1,j+1} &= \tilde{L}_{1j} \cup \{i \in B_1^0 \setminus \tilde{L}_{1j} : g_{1i} < \overline{g_{1j}}\} & \text{if } \gamma_{1j} < \tau_{1j}, \\ \tilde{L}_{1,j+1} &= \tilde{L}_{1j} & \text{otherwise,} \\ \tilde{L}_{2,j+1} &= \tilde{L}_{2j} \cup \{i \in B_2^0 \setminus \tilde{L}_{2j} : g_{2i} < \overline{g_{2j}}\} & \text{if } \gamma_{2j} < \tau_{2j}, \\ \tilde{L}_{2,j+1} &= \tilde{L}_{2j} & \text{otherwise,} \end{aligned} \quad (3.16)$$

and

$$\begin{aligned}\tilde{F}_{1,j+1} &= \tilde{F}_{1_0} \setminus \left(\tilde{U}_{1,j+1} \cup \tilde{L}_{1,j+1} \right), \\ \tilde{F}_{2,j+1} &= \tilde{F}_{2_0} \setminus \left(\tilde{U}_{2,j+1} \cup \tilde{L}_{2,j+1} \right).\end{aligned}\tag{3.17}$$

Lemma 12 *For all $k > j$ in \mathbb{N} , if $\gamma_{1j} \geq \tau_{1j}$, then $\overline{g_{1j}} \geq \overline{g_{1k}}$, and if $\gamma_{1j} < \tau_{1j}$, then $\overline{g_{1j}} \leq \overline{g_{1k}}$. Similarly, for all $k > j$ in \mathbb{N} , if $\gamma_{2j} \geq \tau_{2j}$, then $\overline{g_{2j}} \geq \overline{g_{2k}}$, and if $\gamma_{2j} < \tau_{2j}$, then $\overline{g_{2j}} \leq \overline{g_{2k}}$.*

Proof.

By using the proof of Lemma 1, we can show that for all $k > j$ in \mathbb{N} , if $\gamma_{1j} \geq \tau_{1j}$, then $\overline{g_{1j}} \geq \overline{g_{1k}}$, and if $\gamma_{1j} < \tau_{1j}$, then $\overline{g_{1j}} \leq \overline{g_{1k}}$. Similarly, we can prove that for all $k > j$ in \mathbb{N} if $\gamma_{2j} \geq \tau_{2j}$, then $\overline{g_{2j}} \geq \overline{g_{2k}}$, and if $\gamma_{2j} < \tau_{2j}$, then $\overline{g_{2j}} \leq \overline{g_{2k}}$. \square

Hence, by Lemma 12, for $t = 1, 2$,

$$g_{t_i} > \overline{g_{t_k}} \quad \forall i \in \tilde{U}_{t_i}, \quad \forall k \geq j,\tag{3.18}$$

and

$$g_{t_i} < \overline{g_{t_k}} \quad \forall i \in \tilde{L}_{t_i}, \quad \forall k \geq j.\tag{3.19}$$

Let's stop this process for \mathbf{x}_0 if $\gamma_{1j} = \tau_{1j} = 0$ and $\gamma_{2j} = \tau_{2j} = 0$. Then since B_1^0 and B_1^1 are finite, this must be terminated in finite steps. Let s be the first number such that $\gamma_{1s} = \tau_{1s} = 0$ and $\gamma_{2s} = \tau_{2s} = 0$, then define

$$U_{1_0} = \tilde{U}_{1_s} = \tilde{U}_{1_{s+1}}, \quad L_{1_0} = \tilde{L}_{1_s} = \tilde{L}_{1_{s+1}}, \quad F_{1_0} = \tilde{F}_{1_s} = \tilde{F}_{1_0} \setminus (\tilde{U}_{1_s} \cup \tilde{L}_{1_s}),$$

and

$$U_{2_0} = \tilde{U}_{2_s} = \tilde{U}_{2_{s+1}}, \quad L_{2_0} = \tilde{L}_{2_s} = \tilde{L}_{2_{s+1}}, \quad F_{2_0} = \tilde{F}_{2_s} = \tilde{F}_{2_0} \setminus (\tilde{U}_{2_s} \cup \tilde{L}_{2_s}).$$

Thus, by (3.15), (3.16), (3.18) and (3.19), for $t = 1, 2$,

$$g_{t_k} > \frac{\sum_{i \in \tilde{F}_{t_s}} g_{t_i}}{|\tilde{F}_{t_s}|} = \frac{\sum_{i \in F_{t_0}} g_{t_i}}{|F_{t_0}|} \quad \text{if } k \in U_{t_0} = \tilde{U}_{t_s},\tag{3.20}$$

and

$$g_{t_k} < \frac{\sum_{i \in \tilde{F}_{t_s}} g_{t_i}}{|\tilde{F}_{t_s}|} = \frac{\sum_{i \in F_{t_0}} g_{t_i}}{|F_{t_0}|} \quad \text{if } k \in L_{t_0} = \tilde{L}_{t_s}.\tag{3.21}$$

2.2 Continuing Procedure for Two Partitioned Sets

$$\begin{aligned} & \min \frac{1}{2} [\|\mathbf{x}_1 - \mathbf{y}_1\|^2 + \|\mathbf{x}_2 - \mathbf{y}_2\|^2] \\ & \text{subject to } \mathbf{0} \leq \mathbf{x}_1 \leq \mathbf{1}, \mathbf{0} \leq \mathbf{x}_2 \leq \mathbf{1}, \mathbf{1}^T \mathbf{x}_1 = m = \mathbf{1}^T \mathbf{x}_1 \end{aligned}$$

where $\mathbf{1}^T = (1, 1, \dots, 1)$, $\mathbf{y}_1 = \mathbf{x}_1^k + \alpha g_1$ and $\mathbf{y}_2 = \mathbf{x}_2^k + \alpha g_2$. Let

$$U_{1j} = U_1(\alpha_j) = \{i : x_{1i}(\alpha_j) = 1\}, \quad U_{2j} = U_2(\alpha_j) = \{i : x_{2i}(\alpha_j) = 1\},$$

$$L_{1j} = L_1(\alpha_j) = \{i : x_{1i}(\alpha_j) = 0\}, \quad L_{2j} = L_2(\alpha_j) = \{i : x_{2i}(\alpha_j) = 0\}.$$

Lemma 13 *If $U_1(\alpha_1) = U_1(\alpha_2)$ and $L_1(\alpha_1) = L_1(\alpha_2)$, then for all $\alpha_1 \leq \alpha \leq \alpha_2$,*

$$U_1(\alpha) = U_1(\alpha_1) \quad \text{and} \quad L_1(\alpha) = L_1(\alpha_1).$$

Similarly, if $U_2(\alpha_3) = U_2(\alpha_4)$ and $L_2(\alpha_3) = L_2(\alpha_4)$, then for all $\alpha_3 \leq \alpha \leq \alpha_4$,

$$U_2(\alpha) = U_2(\alpha_3) \quad \text{and} \quad L_2(\alpha) = L_2(\alpha_3).$$

Proof. The proof is the same as the proof of lemma2. □

By Lemma 13, we can conclude that there are only finite number of breaks.

Define $\mathcal{U}_1(\alpha_{j+1})$, $\mathcal{U}_2(\alpha_{j+1})$, $\mathcal{L}_1(\alpha_{j+1})$ and $\mathcal{L}_2(\alpha_{j+1})$ such that

$$\text{if } \mu_t(\alpha_{j+1}) = 0, x_{ti}(\alpha_{j+1}) = 1 \quad \text{then } i \in \mathcal{U}_t(\alpha_{j+1}) \text{ for } t = 1, 2, \quad (3.22)$$

$$\text{if } \nu_t(\alpha_{j+1}) = 0, x_{ti}(\alpha_{j+1}) = 0 \quad \text{then } i \in \mathcal{L}_t(\alpha_{j+1}) \text{ for } t = 1, 2. \quad (3.23)$$

Let

$$F_t^0 = F_t \cup \mathcal{U}_t(\alpha_{j+1}) \cup \mathcal{L}_t(\alpha_{j+1}), \quad U_t^0 = \emptyset, \quad L_t^0 = \emptyset \text{ for } t = 1, 2. \quad (3.24)$$

For $0 \leq l$ and $t = 1, 2$, define

$$\overline{g_t^l} = \frac{\sum_{i \in F_t^l} g_{ti}}{|F_t^l|}, \quad (3.25)$$

$$\gamma_{t_l} = \sum_{i \in \mathcal{U}_t(\alpha_{j+1}) \setminus U_t^l, g_{ti} > \overline{g_t^l}} (g_{ti} - \overline{g_t^l}), \quad (3.26)$$

and

$$\tau_{t_l} = \sum_{i \in \mathcal{L}_t(\alpha_{j+1}) \setminus L_t^l, g_{t_l} < \overline{g}_t^l} (\overline{g}_t^l - g_{t_l}). \quad (3.27)$$

Also, define

$$\begin{aligned} U_t^{l+1} &= U_t^l \cup \left\{ i \in \mathcal{U}_t(\alpha_{j+1}) \setminus U_t^l : g_{t_l} > \overline{g}_t^l \right\} & \text{if } \gamma_{t_l} \geq \tau_{t_l} \\ U_t^{l+1} &= U_t^l & \text{otherwise,} \end{aligned} \quad (3.28)$$

$$\begin{aligned} L_t^{l+1} &= L_t^l \cup \left\{ i \in \mathcal{L}_t(\alpha_{j+1}) \setminus L_t^l : g_{t_l} < \overline{g}_t^l \right\} & \text{if } \gamma_{t_l} < \tau_{t_l} \\ L_t^{l+1} &= L_t^l & \text{otherwise,} \end{aligned} \quad (3.29)$$

and

$$F_t^{l+1} = F_t^0 \setminus (U_t^{l+1} \cup L_t^{l+1}). \quad (3.30)$$

Let us stop this process if $\gamma_{t_l} = \tau_{t_l} = 0, t = 1, 2$. Apply Lemmal2 with $B_t^l = \mathcal{U}_t(\alpha_{j+1}), B_t^0 = \mathcal{L}_t(\alpha_{j+1}), \tilde{F}_{t_0} = F_t^0$ and $\gamma_{t_{j+1}} = \gamma_{t_l}, \tau_{t_{j+1}} = \tau_{t_l}, t = 1, 2$. Then since $\mathcal{U}_t(\alpha_{j+1})$ and $\mathcal{L}_t(\alpha_{j+1})$ for $t = 1, 2$ are finite, this must be stopped in finite steps. Let s be the first number such that $\gamma_{t_s} = \tau_{t_s} = 0, t = 1, 2$, then define $U_{t_{j+1}}$ and $L_{t_{j+1}}$ with $F_{t_{j+1}} = (U_{t_{j+1}} \cup L_{t_{j+1}})^C$ for $t = 1, 2$ as follows :

$$U_{t_{j+1}} = (U_{t_j} \setminus \mathcal{U}_t(\alpha_{j+1})) \cup U_t^s \quad (3.31)$$

and

$$L_{t_{j+1}} = (L_{t_j} \setminus \mathcal{L}_t(\alpha_{j+1})) \cup L_t^s, \quad (3.32)$$

where as similar to (3.20) and (3.21),

$$g_{t_k} > \frac{\sum_{i \in F_{t_{j+1}}} g_{t_i}}{|F_{t_{j+1}}|} \quad \text{if } k \in U_t^s, \quad t = 1, 2, \quad (3.33)$$

and

$$g_{t_k} < \frac{\sum_{i \in F_{t_{j+1}}} g_{t_i}}{|F_{t_{j+1}}|} \quad \text{if } k \in L_t^s, \quad t = 1, 2. \quad (3.34)$$

Now we define functions $\mathbf{x}(F_j, \alpha_j), \mu(F_j, \alpha_j), \lambda(F_j, \alpha_j)$, with $0 = \alpha_0 < \alpha_1 < \alpha_2 \cdots$, where $(F_{t_{j+1}}, \alpha_{j+1})$ are obtained from $(F_{t_j}, \alpha_j), t = 1, 2$ in the following way;

For $t = 1, 2$,

$$\lambda_t(\alpha) = \frac{(|U_{t_j}| - m + \sum_{i \in F_{t_j}} y_{t_i}(\alpha))}{|F_{t_j}|}, \text{ if } \alpha_j \leq \alpha \leq \alpha_{j+1} \quad (3.35)$$

$$= \frac{|U_{t_j}| - m}{|F_{t_j}|} + \frac{\sum_{i \in F_{t_j}} x_{t_{0_i}}}{|F_{t_j}|} + \frac{\alpha \sum_{i \in F_{t_j}} g_{t_i}}{|F_{t_j}|},$$

$$x_{t_i}(\alpha) = y_{t_i}(\alpha) - \lambda(F_{t_j}, \alpha), \text{ if } i \in F_{t_j} \text{ and } \alpha_j \leq \alpha \leq \alpha_{j+1} \quad (3.36)$$

$$= x_{t_{0_i}} + \alpha \left(g_{t_i} - \frac{\sum_{i \in F_{t_j}} g_{t_i}}{|F_{t_j}|} \right) - \frac{|U_{t_j}| - m}{|F_{t_j}|} - \frac{\sum_{i \in F_{t_j}} x_{t_{0_i}}}{|F_{t_j}|},$$

$$\mu_i(\alpha) = y_{t_i}(\alpha) - 1 - \lambda(F_{t_j}, \alpha), \text{ if } i \in U_{t_j} \text{ and } \alpha_j \leq \alpha \leq \alpha_{j+1} \quad (3.37)$$

$$= x_{t_{0_i}} - 1 + \alpha \left(g_{t_i} - \frac{\sum_{i \in F_{t_j}} g_{t_i}}{|F_{t_j}|} \right) - \frac{|U_{t_j}| - m}{|F_{t_j}|} - \frac{\sum_{i \in F_{t_j}} x_{t_{0_i}}}{|F_{t_j}|},$$

$$\nu_{t_i}(\alpha) = \lambda(F_{t_j}, \alpha) - y_{t_i}(\alpha), \text{ if } i \in L_{t_j} \text{ and } \alpha_j \leq \alpha \leq \alpha_{j+1} \quad (3.38)$$

$$= \frac{|U_{t_j}| - m}{|F_{t_j}|} + \frac{\sum_{i \in F_{t_j}} x_{t_{0_i}}}{|F_{t_j}|} + \alpha \left(\frac{\sum_{i \in F_{t_j}} g_{t_i}}{|F_{t_j}|} - g_{t_i} \right) - x_{t_{0_i}}$$

and

$$\alpha_{j+1} = \min \left\{ \begin{array}{l} \sup\{\alpha : \mu_{1i}(\alpha) > 0 \ \forall i \in U_{1j}, \mu_{1i}(\alpha_j^+) \neq 0\}, \\ \sup\{\alpha : \nu_{1i}(\alpha) > 0 \ \forall i \in L_{1j}, \nu_{1i}(\alpha_j^+) \neq 0\}, \\ \sup\{\alpha : 0 < x_{1i}(F_{1j}, \alpha) < 1 \ \forall i \in F_{1j}, x_{1i}(\alpha_j^+) \neq 0, 1\}, \\ \sup\{\alpha : \mu_{2i}(\alpha) > 0 \ \forall i \in U_{2j}, \mu_{2i}(\alpha_j^+) \neq 0\}, \\ \sup\{\alpha : \nu_{2i}(\alpha) > 0 \ \forall i \in L_{2j}, \nu_{2i}(\alpha_j^+) \neq 0\}, \\ \sup\{\alpha : 0 < x_{2i}(F_{2j}, \alpha) < 1 \ \forall i \in F_{2j}, x_{2i}(\alpha_j^+) \neq 0, 1\} \end{array} \right\}. \quad (3.39)$$

By (3.2), (3.36), (3.37), (3.38) and (3.39), for $t = 1, 2$, we have

$$\mu_{t_i}(\alpha) = 0 \quad \text{for } i \notin U_{t_j}, \alpha_j \leq \alpha \leq \alpha_{j+1}, \quad (3.40)$$

$$\nu_{t_i}(\alpha) = 0 \quad \text{for } i \notin L_{t_j}, \alpha_j \leq \alpha \leq \alpha_{j+1}, \quad (3.41)$$

$$x_{t_i}(\alpha) = 1 \quad \text{for } i \in U_{t_j}, \alpha_j \leq \alpha \leq \alpha_{j+1} \quad (3.42)$$

and

$$x_{t_i}(\alpha) = 0 \quad \text{for } i \in L_{t_j}, \alpha_j \leq \alpha \leq \alpha_{j+1}. \quad (3.43)$$

Theorem 14 For all i and α , $x_{t_i}(\alpha)$, $\lambda_t(\alpha)$, $\mu_{t_i}(\alpha)$ and $\nu_{t_i}(\alpha)$ are continuous piece-wise linear functions of α , and $0 \leq x_{t_i}(\alpha) \leq 1$ for $t = 1, 2$, that is $0 \leq \mathbf{x}_1 \leq 1$ and $0 \leq \mathbf{x}_2 \leq 1$.

This theorem is followed by theorem 3.

Hence, the condition (3.6) is satisfied. Now, if we hold the condition (3.5), $\mu_t \geq \mathbf{0}$, $\nu_t \geq \mathbf{0}$ for $t = 1, 2$, then $\mathbf{x}(\alpha) = \begin{pmatrix} \mathbf{x}_1(\alpha) \\ \mathbf{x}_2(\alpha) \end{pmatrix}$ is the solution of (3.1). Let $F_{t_I} = (B_t^1 \cup B_t^0)^C$ where $B_t^1 = \{i : x_{t_0_i} = 1\}$, $B_t^0 = \{i : x_{t_0_i} = 0\}$, $t = 1, 2$. Then by the same argument as before in previous chapter, we can show the condition (3.5) is hold. Therefore $\mathbf{x}(\alpha) = \begin{pmatrix} \mathbf{x}_1(\alpha) \\ \mathbf{x}_2(\alpha) \end{pmatrix}$ is the unique minimizer of (3.1).

We have the same argument for extreme point of convex set and stopping criteria for a given search direction as before.

CHAPTER 4 ALGORITHM

We will implement algorithm to obtain a local minimizer of the quadratic programming problem

$$\begin{aligned} & \min f(\mathbf{x}), \quad f(\mathbf{x}) = (\mathbf{1} - \mathbf{x})^T \mathbf{A} \mathbf{x} \\ & \text{subject to } \mathbf{0} \leq \mathbf{x} \leq \mathbf{1}, \mathbf{1}^T \mathbf{x} = m \end{aligned}$$

by using gradient projection method with the step-size α^{\min} such that

$$f(\mathbf{x}(\alpha^{\min})) = \min_{\alpha \geq 0} f(\mathbf{x}(\alpha))$$

where $\mathbf{x}(\alpha) = Proj_K[\mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k)]$.

Let U, L be upper and lower active sets of indices at current iteration, respectively. $F = (U \cup L)^C$. That is, F : free set, inactive set.

Let $\mathbf{g}^k : -\nabla f(\mathbf{x}^k) = -(\mathbf{A}\mathbf{1} - 2\mathbf{A}\mathbf{x}^k)$, $\bar{g} = \frac{\sum_{i \in F} g_i}{|F|}$ and $\bar{c} = \frac{|U| - m + \sum_{i \in F} x_{0_i}}{|F|}$ where $\bar{c} = 0$ at the starting point. Let $\mathbf{1}_{SF}$, \mathbf{g}_{SF} and $\mathbf{x}_{0_{SF}}$ are vectors which have length of n with 1, g_i and x_{0_i} respectively for all components for $i \in F$ and 0 for others. $\mathbf{1}_{SU}$ and $\mathbf{1}_{SL}$ are vectors which have length of n with 1 for all components for $i \in U$, $i \in L$ respectively and 0 for others.

$$\begin{aligned} f'(\mathbf{x}(\alpha)) &= \frac{df(\mathbf{x}(\alpha))}{d\alpha} = \frac{d\mathbf{x}}{d\alpha} \frac{df(\mathbf{x}(\alpha))}{d\mathbf{x}} \\ &= (\mathbf{g}_F - \bar{g}\mathbf{1}_F)^T [\mathbf{A}_F \mathbf{1} - 2\mathbf{A}_F \mathbf{x}(\alpha)] \\ &= (\mathbf{g}_F - \bar{g}\mathbf{1}_F)^T [\mathbf{A}_{FF} \mathbf{1}_F + \mathbf{A}_{FU} \mathbf{1}_U + \mathbf{A}_{FL} \mathbf{1}_L - 2\mathbf{A}_{FF} \mathbf{x}_F(\alpha) - 2\mathbf{A}_{FU} \mathbf{1}_U] \\ &= (\mathbf{g}_F - \bar{g}\mathbf{1}_F)^T [\mathbf{A}_{FF} \mathbf{1}_F - 2\mathbf{A}_{FF} \mathbf{x}_F(\alpha) - \mathbf{A}_{FU} \mathbf{1}_U + \mathbf{A}_{FL} \mathbf{1}_L] \\ &= \mathbf{g}_F^T \mathbf{A}_F [(1 + 2\alpha\bar{g} + 2\bar{c})\mathbf{1}_{SF} - 2(\alpha\mathbf{g}_{SF} + \mathbf{x}_{0_{SF}}) + (\mathbf{1}_{SL} - \mathbf{1}_{SU})] \\ &\quad - \mathbf{1}_F^T \mathbf{A}_F \bar{g} [(1 + 2\alpha\bar{g} + 2\bar{c})\mathbf{1}_{SF} - 2(\alpha\mathbf{g}_{SF} + \mathbf{x}_{0_{SF}}) + (\mathbf{1}_{SL} - \mathbf{1}_{SU})]. \end{aligned}$$

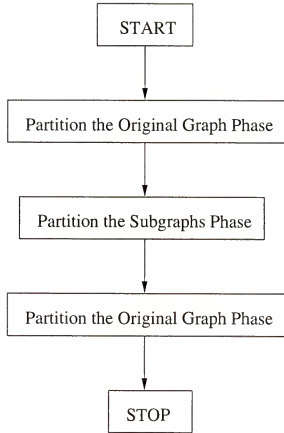


Figure 4.1: Outline of program

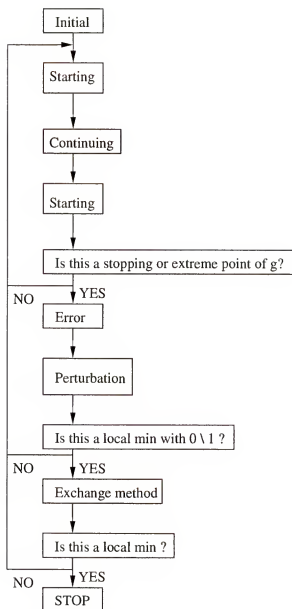


Figure 4.2: Outline of partition of the original graph

$$\begin{aligned}
f''(\mathbf{x}(\alpha)) &= \frac{d\mathbf{x}}{d\alpha} \frac{d^2 f(\mathbf{x}(\alpha))}{d\alpha^2} \left(\frac{d\mathbf{x}}{d\alpha} \right)^T \\
&= (\mathbf{g}_F - \bar{\mathbf{g}}\mathbf{1}_F)^T [-2\mathbf{A}_{FF}] (\mathbf{g}_F - \bar{\mathbf{g}}\mathbf{1}_F) \\
&= -2(\mathbf{g}_F^T \mathbf{A}_{FF} - \bar{\mathbf{g}}\mathbf{1}_F^T \mathbf{A}_{FF})(\mathbf{g}_F - \bar{\mathbf{g}}\mathbf{1}_F).
\end{aligned}$$

1 Initial Point

We can get an initial point in the following three ways :

- (i) \mathbf{x}_c , the centroid of the convex set, $K = \{\mathbf{0} \leq \mathbf{x} \leq \mathbf{1}, \mathbf{1}^T \mathbf{x} = m\}$
- (ii) The point which has the best function value of the ball centered at \mathbf{x}_c in K .
- (iii) $\mathbf{x}_c + \mathbf{y}$, the projection onto K of the point whose function value is the best of the ball which contains K .

1.1 The Solution of the Ball Centered at \mathbf{x}_c in K

Let $\mathbf{x}_c + \mathbf{y}$ be the solution of

$$\begin{aligned}
&\min (\mathbf{x}_c + \mathbf{y})^T \mathbf{A} (\mathbf{1} - \mathbf{x}_c - \mathbf{y}) \\
&\text{subject to } \|\mathbf{y}\| \leq r, \mathbf{1}^T \mathbf{y} = 0
\end{aligned} \tag{4.1}$$

where \mathbf{x}_c is the centroid of K , $\frac{m}{n}\mathbf{1}$ and

$$\begin{aligned}
r &= (\|\mathbf{x}_c - \text{the closest point on the boundary of } K \text{ from } \mathbf{x}_c\|) \\
&= \min \left[\sqrt{2} \frac{m}{n}, \sqrt{2} \left(1 - \frac{m}{n} \right) \right].
\end{aligned}$$

Then the radius, r is the minimum distance from centroid to boundary of K and so, $0 \leq (\mathbf{x}_c + \mathbf{y})_i \leq 1$ for all i . Since

$$(\mathbf{x}_c + \mathbf{y})^T \mathbf{A} (\mathbf{1} - \mathbf{x}_c - \mathbf{y}) - \mathbf{x}_c^T \mathbf{A} (\mathbf{1} - \mathbf{x}_c) = -\mathbf{y}^T \mathbf{A} \mathbf{y} + \mathbf{y}^T \mathbf{A} (\mathbf{1} - 2\mathbf{x}_c),$$

the solutions for (4.1) is equal to the solution of the following :

$$\begin{aligned}
&\min -\mathbf{y}^T \mathbf{A} \mathbf{y} + \mathbf{y}^T \mathbf{A} (\mathbf{1} - 2\mathbf{x}_c) \\
&\text{subject to } \|\mathbf{y}\| \leq r, \mathbf{1}^T \mathbf{y} = 0
\end{aligned} \tag{4.2}$$

where $\mathbf{1} - 2\mathbf{x}_c = (1 - \frac{2m}{n})\mathbf{1}$.

Let $\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_n$ be orthonormal eigenvectors corresponding to the eigenvalues, $\lambda_1, \lambda_2, \dots, \lambda_n$ of \mathbf{PAP} where $\mathbf{P} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T$. Take $\mathbf{f}_1 = \frac{1}{\sqrt{n}}\mathbf{1}$, $\lambda_1 = 0$ and $\mathbf{y} = \sum_{i=2}^n c_i \mathbf{f}_i$. Then

$$\mathbf{PAP}\mathbf{f}_i = \lambda_i \mathbf{f}_i$$

and since $\mathbf{1}^T \mathbf{y} = 0$, $\mathbf{P}\mathbf{y} = \mathbf{y}$. So,

$$\begin{aligned} \mathbf{y}^T \mathbf{A} \mathbf{y} &= \mathbf{y}^T \mathbf{PAP} \mathbf{y} \\ &= \mathbf{y}^T \sum_{i=2}^n \mathbf{PAP}(c_i \mathbf{f}_i) \\ &= \left(\sum_{i=2}^n c_i \mathbf{f}_i \right)^T \sum_{i=2}^n c_i \lambda_i \mathbf{f}_i \\ &= \sum_{i=2}^n c_i^2 \lambda_i. \end{aligned}$$

Also,

$$\begin{aligned} \mathbf{y}^T \mathbf{A} (\mathbf{1} - 2\mathbf{x}_c) &= \left(\sum_{i=2}^n c_i \mathbf{f}_i \right)^T \mathbf{A} \left(1 - \frac{2m}{n} \right) \mathbf{1} \\ &= \sum_{i=2}^n c_i \left(1 - \frac{2m}{n} \right) \mathbf{f}_i^T \mathbf{A} \mathbf{1}. \end{aligned}$$

Hence,

$$(4.2) \equiv \min_{\text{subject to } \|\mathbf{y}\| \leq r, \mathbf{1}^T \mathbf{y} = 0} -\mathbf{y}^T \mathbf{PAP} \mathbf{y} + \mathbf{y}^T \mathbf{PA} (\mathbf{1} - 2\mathbf{x}_c) \quad (4.3)$$

and

$$(4.3) \equiv \min_{\text{subject to } \sum_{i=2}^n (-c_i^2 \lambda_i + 2d_i c_i) \leq r^2, \mathbf{1}^T \sum_{i=2}^n c_i \mathbf{f}_i = 0} \sum_{i=2}^n (-c_i^2 \lambda_i + 2d_i c_i) \quad (4.4)$$

where $d_i = \frac{1}{2} (1 - \frac{2m}{n}) \mathbf{f}_i^T \mathbf{PA} \mathbf{1}$. If $\mathbf{P}\mathbf{y} = \mathbf{y}$, then $\mathbf{1}^T \mathbf{y} = 0$. Hence the solution for (4.2) is equal to the solution of the following problem :

$$\min_{\text{subject to } \frac{1}{2} \sum_{i=2}^n (-c_i^2 \lambda_i + 2d_i c_i) \leq r^2, \mathbf{1}^T \sum_{i=2}^n c_i \mathbf{f}_i = 0} \sum_{i=2}^n (-c_i^2 \lambda_i + 2d_i c_i) \quad (4.5)$$

Lemma 15 *If \mathbf{PAP} has a positive eigenvalue, then $\sum_{i=2}^n c_i^2 = r^2$ at the minimum of (4.5).*

Let $x_i = 1, x_j = -1$ and $x_k = 0$ for all $k \neq i, j$. Then $\mathbf{P}\mathbf{x} = \mathbf{x}$.

$$\mathbf{x}^T \mathbf{P} \mathbf{A} \mathbf{P} \mathbf{x} = (\mathbf{P}\mathbf{x})^T \mathbf{A} (\mathbf{P}\mathbf{x}) = \mathbf{x}^T \mathbf{A} \mathbf{x}.$$

If $a_{ij} = 0$ for some i, j , then $a_{ii} + a_{jj} = 2 > 0$. Thus $\mathbf{P}\mathbf{A}\mathbf{P}$ has a positive eigenvalue.

Since $\mathbf{P}\mathbf{A}\mathbf{P}$ has a positive eigenvalue, by the lemma 15,

$$(4.5) \equiv \begin{array}{l} \min \frac{1}{2} \sum_{i=2}^n (-c_i^2 \lambda_i + 2d_i c_i) \\ \text{subject to } \frac{1}{2} \sum_{i=2}^n c_i^2 = \frac{r^2}{2}. \end{array} \quad (4.6)$$

Let us see the form of a solution to (4.6) by first and second order optimality conditions. At minimum, let μ be a multiplier for constraint, then

$$\begin{aligned} -c_i \lambda_i + d_i + \mu c_i &= 0 \\ \Rightarrow c_i &= \frac{-d_i}{\mu - \lambda_i} \\ \Rightarrow \sum_{i=2}^n \frac{d_i^2}{(\mu - \lambda_i)^2} &= r^2. \end{aligned}$$

Since $\frac{1}{2} \sum_{i=2}^n (-c_i^2 \lambda_i + 2d_i c_i)$ and $\frac{1}{2} \sum_{i=2}^n c_i^2$ are in C^2 , we can use the second order optimality condition. Hence if \mathbf{c} is a local minimizer of (4.6), then for all $\mathbf{y} \in \mathbb{R}^n$ such that $\sum_{i=2}^n c_i y_i = 0$,

$$\mathbf{y}^T \left(\nabla^2 \left(\frac{1}{2} \sum_{i=2}^n (-c_i^2 \lambda_i + 2d_i c_i) \right) + \mu \nabla^2 \left(\frac{1}{2} \sum_{i=2}^n c_i^2 \right) \right) \mathbf{y} \geq 0. \quad (4.7)$$

But $(\nabla^2(\frac{1}{2} \sum_{i=2}^n (-c_i^2 \lambda_i + 2d_i c_i)) + \mu \nabla^2(\frac{1}{2} \sum_{i=2}^n c_i^2))$ is a matrix whose diagonal is $(0, \mu - \lambda_2, \mu - \lambda_3, \dots, \mu - \lambda_n)$ and all the other entries are zeroes. Hence, (4.7) is equivalent to

$$\sum_{i=2}^n (\mu - \lambda_i) y_i^2 \geq 0. \quad (4.8)$$

Thus, by the second order optimality conditions, we know that the solution to (4.6) can be expressed in the following way: Let μ be the largest number such that

$$\sum_{i=2}^n \frac{d_i^2}{(\mu - \lambda_i)^2} = r^2. \quad (4.9)$$

$$(4.3) \equiv \begin{array}{l} \min -\mathbf{y}^T \mathbf{P} \mathbf{A} \mathbf{P} \mathbf{y} + \mathbf{y}^T \mathbf{P} \mathbf{A} \left(1 - \frac{2m}{n}\right) \mathbf{1} \\ \text{subject to } \mathbf{y}^T \mathbf{y} = r^2, \mathbf{1}^T \mathbf{y} = 0. \end{array} \quad (4.10)$$

Hence, by using the projection onto $\mathbf{1}$, $\mathbf{P} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ and the first and second order optimality conditions, we know that the minimum occurs at the boundary and the form of the solution of (4.2) is (4.9).

Now, we will find a solution in \mathbb{R}^n of (4.2). We can find a unit vector \mathbf{v}_1 which is orthogonal to $\mathbf{1}$ from a random vector on the unit sphere in \mathbb{R}^{n-1} by using Householder transformation. Let $\mathbf{v}_1, \dots, \mathbf{v}_k$ be orthonormal vectors for which $\mathbf{1}^T \mathbf{v}_i = 0$, $i = 1, 2, \dots, k$ and $\mathbf{y} = \mathbf{V}\mathbf{z}$ where \mathbf{V} is the matrix with columns of \mathbf{v}_i 's. Then since $\mathbf{z}^T \mathbf{z} = \mathbf{z}^T \mathbf{V}^T \mathbf{V} \mathbf{z} = \mathbf{y}^T \mathbf{y}$ and $\mathbf{1}^T \mathbf{y} = 0$, we have

$$(4.3) \equiv \min_{\text{subject to } \mathbf{z}^T \mathbf{z} = r^2} (-\mathbf{z}^T \mathbf{V}^T \mathbf{P} \mathbf{A} \mathbf{P} \mathbf{V} \mathbf{z} + \mathbf{z}^T \mathbf{V}^T \mathbf{f}) \quad (4.11)$$

where $\mathbf{f} = \mathbf{P} \mathbf{A} \left(1 - \frac{2m}{n}\right) \mathbf{1}$. Let $\mathbf{P} \mathbf{A} \mathbf{P} = \mathbf{S} - \mathbf{T}_1$ where \mathbf{S} is invertible and $\mathbf{M} = \mathbf{S}^{-1} \mathbf{T}_1$. If \mathbf{V}_k denotes the matrix whose columns are $\{\mathbf{v}_1, \dots, \mathbf{v}_k\}$ and \mathbf{H} is the $(k+1) \times k$ upper Hessenberg matrix whose elements are given by the Arnoldi process, then

$$\mathbf{M} \mathbf{V}_k = \mathbf{V}_{k+1} \mathbf{H}.$$

Let $\mathbf{S} = -w\mathbf{I}$, then since $\mathbf{T}_1 = \mathbf{S} - \mathbf{P} \mathbf{A} \mathbf{P}$,

$$\mathbf{M} = \mathbf{I} + \frac{1}{w} \mathbf{P} \mathbf{A} \mathbf{P}.$$

Let $w > \rho(\mathbf{P} \mathbf{A} \mathbf{P})$. Then $w > -$ (smallest eigenvalue of $\mathbf{P} \mathbf{A} \mathbf{P}$) and so all eigenvalues of \mathbf{M} are positive. Actually, since our matrix \mathbf{M} is symmetric, \mathbf{H} is the $(k+1) \times k$ tridiagonal matrix (see [8]).

$$\begin{aligned} \mathbf{V}^T \mathbf{P} \mathbf{A} \mathbf{P} \mathbf{V} &= \mathbf{V}^T (\mathbf{S} - \mathbf{T}_1) \mathbf{V} \\ &= \mathbf{V}^T \mathbf{S} (\mathbf{I} - \mathbf{M}) \mathbf{V} \\ &= \mathbf{S} \mathbf{V}_k^T (\mathbf{V}_k - \mathbf{V}_{k+1} \mathbf{H}) \\ &= \mathbf{S} (\mathbf{I} - \mathbf{T}_2) \\ &= -w (\mathbf{I} - \mathbf{T}_2) \end{aligned}$$

where \mathbf{T}_2 is tridiagonal since \mathbf{H} is tridiagonal.

1. Generate a random vector $\bar{\mathbf{x}}$ on the unit sphere in \mathbb{R}^{n-1} .
2. Generate a unit vector \mathbf{v}_1 with $\mathbf{1}^T \mathbf{v}_1 = 0$ from $\bar{\mathbf{x}}$.
3. Use Arnoldi algorithm to generate \mathbf{V} and \mathbf{T}_2 with above \mathbf{v}_1 .
4. Use the following algorithms(Newton And Bisection) to solve

$$(4.11) \equiv \begin{array}{l} \min -(\mathbf{z}^T \mathbf{T}_3 \mathbf{z}) + \mathbf{t}^T \mathbf{z}, \quad \mathbf{t} = \mathbf{V}^T \mathbf{f} \\ \text{subject to } \mathbf{z}^T \mathbf{z} = r^2 \end{array} \quad (4.12)$$

where \mathbf{T}_3 is $-\mathbf{w}(\mathbf{I} - \mathbf{T}_2)$ above.

5. $\mathbf{y} = \mathbf{V} \mathbf{z}$ (This will be the solution of (4.2)).

We will use Arnoldi process to choose \mathbf{V} to be a matrix whose columns are orthonormal and $\mathbf{V}^T \mathbf{P} \mathbf{A} \mathbf{P} \mathbf{V}$ is tridiagonal. If the starting vector for the Arnoldi process has components summing to zero, and if we use

$$\mathbf{M} = \mathbf{I} + \frac{1}{w} \mathbf{P} \mathbf{A} \mathbf{P},$$

then all the succeeding vectors satisfy this condition (summing the components is zero). At first, generate a random vector $\bar{\mathbf{x}}$ on the unit sphere in \mathbb{R}^{n-1} by using Gaussian distribution with *mean* = 0 (see [8]). Let

$$\bar{\mathbf{x}} = (x_1 \cdots x_{n-1})^T \text{ and } \bar{\bar{\mathbf{x}}} = \begin{pmatrix} \bar{\mathbf{x}} \\ 0 \end{pmatrix}.$$

Now choose $\hat{\mathbf{H}} = \mathbf{I} - 2\mathbf{w}\mathbf{w}^T$ such that

$$\hat{\mathbf{H}} \mathbf{1} = \sqrt{n} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}$$

Since $\hat{\mathbf{H}}$ is symmetric, the last row and column of $\hat{\mathbf{H}}$ are $\frac{1}{\sqrt{n}} \mathbf{1}^T$ and $\frac{1}{\sqrt{n}} \mathbf{1}$. $\|\hat{\mathbf{H}} \bar{\bar{\mathbf{x}}}\| = 1$ since $\bar{\bar{\mathbf{x}}}$ is a unit vector. $\hat{\mathbf{H}}$ rotates $\mathbf{1}$ to the n -th axis, and when a vector on the n -th axis is multiplied by $\hat{\mathbf{H}}$, it is rotated to $\mathbf{1}$. By orthogonality, a vector which is orthogonal to n -th coordinate axis is mapped into a vector orthogonal to $\mathbf{1}$. Let

$\mathbf{a} = \sqrt{n}(0 \ \cdots \ 0 \ 1)^T$. Then

$$\begin{aligned}\mathbf{w} &= \frac{\mathbf{1} - \mathbf{a}}{\|\mathbf{1} - \mathbf{a}\|_2} \\ &= \frac{1}{\sqrt{(n-1) + (1 - \sqrt{n})^2}}(1 \ \cdots \ 1 \ 1 - \sqrt{n})^T.\end{aligned}$$

Hence,

$$\begin{aligned}\hat{\mathbf{H}} &= \mathbf{I} - 2\mathbf{w}\mathbf{w}^T \\ &= \mathbf{I} - \frac{1}{n - \sqrt{n}}(1 \ \cdots \ 1 \ 1 - \sqrt{n})^T(1 \ \cdots \ 1 \ 1 - \sqrt{n}),\end{aligned}$$

and

$$\begin{aligned}\hat{\mathbf{H}}\bar{\mathbf{x}} &= \hat{\mathbf{H}} \begin{pmatrix} \bar{\mathbf{x}} \\ 0 \end{pmatrix} \\ &= \left(\mathbf{I} - \frac{1}{n - \sqrt{n}} \begin{pmatrix} 1 \\ \vdots \\ 1 \\ 1 - \sqrt{n} \end{pmatrix} (1 \ \cdots \ 1 \ 1 - \sqrt{n}) \right) \begin{pmatrix} x_1 \\ \vdots \\ x_{n-1} \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} \bar{\mathbf{x}} \\ 0 \end{pmatrix} - \frac{\bar{\mathbf{x}}^T \mathbf{1}}{n - \sqrt{n}} \begin{pmatrix} \mathbf{1} \\ 1 - \sqrt{n} \end{pmatrix} \\ &= \begin{pmatrix} \bar{\mathbf{x}} - \frac{\bar{\mathbf{x}}^T \mathbf{1}}{n - \sqrt{n}} \mathbf{1} \\ \frac{\bar{\mathbf{x}}^T \mathbf{1}}{\sqrt{n}} \end{pmatrix}.\end{aligned}$$

Let $\mathbf{v}_1 = \begin{pmatrix} \bar{\mathbf{x}} - \frac{\bar{\mathbf{x}}^T \mathbf{1}}{n - \sqrt{n}} \mathbf{1} \\ \frac{\bar{\mathbf{x}}^T \mathbf{1}}{\sqrt{n}} \end{pmatrix}$. Then \mathbf{v}_1 is a unit vector which is orthogonal to $\mathbf{1}$.

Algorithm (Arnoldi for symmetric version)

```

 $\mathbf{v}_1$  is given
for  $j = 1 : k - 1$ 
     $\mathbf{s} \leftarrow \mathbf{M}\mathbf{v}_j$ 
    for  $i = j - 1 : j$ 
         $h_{ij} \leftarrow \mathbf{v}_i^T \mathbf{s} \quad (= \mathbf{v}_i^T \mathbf{M}\mathbf{v}_j)$ 
     $\mathbf{s} \leftarrow \mathbf{s} - h_{ij} \mathbf{v}_i$ 
end
 $h_{j+1,j} \leftarrow \|\mathbf{s}\|$ 
if  $j_{j+1,j} = 0$ ,
    stop
end
 $\mathbf{v}_{j+1} \leftarrow \mathbf{s} / h_{j+1,j}$ 
end
```

end Algorithm

By Arnoldi process, we generate \mathbf{V} and \mathbf{T}_2 . Then we can get \mathbf{T}_3 and \mathbf{t} in (4.12). Now we are ready to solve (4.12).

Let us consider the problem of the form

$$\begin{aligned} & \min\{\psi(w) : \|w\| \leq r\} \\ \psi(w) &= \frac{1}{2}w^T Bw + \mathbf{g}^T w, \end{aligned} \quad (4.13)$$

where r is a positive parameter, $\|\cdot\|$ is the Euclidean norm in \mathbb{R}^n , and with $\mathbf{g} \in \mathbb{R}^n$, and $B \in \mathbb{R}^{n \times n}$ a symmetric matrix.

Lemma 16 *If \mathbf{p} is a solution to (4.13) then \mathbf{p} is a solution to an equation of the form*

$$(B + \lambda I)\mathbf{p} = -\mathbf{g}, \quad (4.14)$$

with $B + \lambda I$ positive semidefinite, $\lambda \geq 0$, and $\lambda(r - \|\mathbf{p}\|) = 0$.

Lemma 17 *Let $\lambda \in \mathbb{R}$, $\mathbf{p} \in \mathbb{R}^n$ satisfy (4.14) with $B + \lambda I$ positive semidefinite.*

1. *If $\lambda = 0$ and $\|\mathbf{p}\| < r$ then \mathbf{p} solves (4.13).*
2. *\mathbf{p} solves $\psi(\mathbf{p}) = \min\{\psi(w) : \|w\| = \|\mathbf{p}\|\}$.*
3. *If $\lambda \geq 0$ and $\|\mathbf{p}\| = r$ then \mathbf{p} solves (4.13).*

If $B + \lambda I$ is positive definite, then \mathbf{p} is the only solution to (4.13).

See [19] for the proof of Lemma 16 and Lemma 17. Note that these lemmas provide necessary and sufficient conditions for a point $\mathbf{p} \in \mathbb{R}^n$ to be a solution to (4.13) and that there is no gap between the necessary and sufficient conditions.

$$(4.12) \equiv \begin{aligned} & \min \frac{1}{2}\mathbf{z}^T(-\mathbf{T})\mathbf{z} + \mathbf{t}^T\mathbf{z} \\ & \text{subject to } \mathbf{z}^T\mathbf{z} = r^2. \end{aligned} \quad (4.15)$$

where $\mathbf{T} = 2\mathbf{T}_3 = -2w(\mathbf{I} - \mathbf{T}_2)$. Let \mathbf{p} be the solution of (4.15), then $\mathbf{p}^T \mathbf{p} = r^2$. Hence, if we have $-\mathbf{T} + \lambda \mathbf{I}$ positive definite with $\lambda \geq 0$ then $\mathbf{p} = -(-\mathbf{T} + \lambda \mathbf{I})^{-1} \mathbf{t}$ is the solution to (4.12). But if $\mathbf{t} = \mathbf{0}$, then it is not true. So, we can use this only if $\mathbf{t} \neq \mathbf{0}$. We will discuss $\mathbf{t} = \mathbf{0}$ case later.

Assume $\mathbf{t} \neq \mathbf{0}$. Let λ_{\max} be the largest eigenvalue of \mathbf{T} . If $\lambda > \lambda_{\max}$, then $-\mathbf{T} + \lambda \mathbf{I}$ is positive definite. We can evaluate the upper bound of the eigenvalues of \mathbf{T} by Gershgorin's theorem. Let

$$\mathbf{p}_\alpha = -(-\mathbf{T} + \alpha \mathbf{I})^{-1} \mathbf{t}.$$

We know that the function $\|\mathbf{p}_\alpha\|^2$ is a rational function in α with second order poles on a subset of the positives of the eigenvalues of the symmetric matrix \mathbf{T} as (4.9). The rational structure of $\|\mathbf{p}_\alpha\|^2$ may be exploited by applying Newton's method to the zero finding problem

$$\phi(\alpha) \equiv \|\mathbf{p}_\alpha\| - r = 0. \quad (4.16)$$

If $\phi(\alpha)$ is positive, Newton's method is very efficient when applied to (4.16) since ϕ is monotone decreasing on (λ_{\max}, ∞) . Moreover, the computation of the Cholesky factorization of $-\mathbf{T} + \alpha \mathbf{I}$ makes it possible to compute the necessary derivative whenever $\alpha \in (\lambda_{\max}, \infty)$. Let $\mathbf{R}^T \mathbf{R}$ is the Cholesky factorization of $-\mathbf{T} + \lambda \mathbf{I}$ with $\mathbf{R} \in \mathbb{R}^{n \times n}$ upper triangular and \mathbf{p} be the solution of $\mathbf{R}^T \mathbf{R} \mathbf{p} = -\mathbf{t}$.

$$\begin{aligned} \phi'(\alpha) = \|\mathbf{p}\|' &= \frac{\mathbf{p}^T \mathbf{p}'}{\|\mathbf{p}\|} \\ &= \frac{1}{\|\mathbf{p}\|} \left(-(-\mathbf{T} + \alpha \mathbf{I})^{-1} \mathbf{t} \right)^T (-\mathbf{T} + \alpha \mathbf{I})^{-1} \mathbf{I} (-\mathbf{T} + \alpha \mathbf{I})^{-1} \mathbf{t} \\ &= -\frac{1}{\|\mathbf{p}\|} (\mathbf{R}^{-1} \mathbf{R}^{-T} \mathbf{t})^T \mathbf{R}^{-1} \mathbf{R}^{-T} \mathbf{R}^{-1} \mathbf{R}^{-T} \mathbf{t} \\ &= -\frac{1}{\|\mathbf{p}\|} \mathbf{t}^T \mathbf{R}^{-1} \mathbf{R}^{-T} \mathbf{R}^{-1} \mathbf{R}^{-T} \mathbf{R}^{-1} \mathbf{R}^{-T} \mathbf{t} \\ &= -\frac{1}{\|\mathbf{p}\|} \mathbf{p}^T \mathbf{R}^{-1} \mathbf{R}^{-T} \mathbf{p} \\ &= -\frac{1}{\|\mathbf{p}\|} \|\mathbf{R}^{-T} \mathbf{p}\|^2 \end{aligned}$$

$$= -\frac{1}{\|\mathbf{p}\|} \|\mathbf{q}\|^2$$

where $\mathbf{R}^T \mathbf{q} = \mathbf{p}$. Therefore,

$$\frac{\phi(\lambda)}{\phi'(\lambda)} = -\frac{\|\mathbf{p}\|(\|\mathbf{p}\| - r)}{\|\mathbf{q}\|^2}.$$

Hence, without computing the eigensystem of \mathbf{T} , the following algorithm (Algorithm 3.2 in [19]) updates λ by Newton's method applied to (4.16).

Algorithm (Newton) Let $\lambda \geq 0$ with $-\mathbf{T} + \lambda \mathbf{I}$ positive definite and $r > 0$ be given.

1. Factor $-\mathbf{T} + \lambda \mathbf{I} = \mathbf{R}^T \mathbf{R}$;
2. Solve $\mathbf{R}^T \mathbf{R} \mathbf{p} = -\mathbf{t}$;
3. Solve $\mathbf{R}^T \mathbf{q} = \mathbf{p}$;
4. Let $\lambda = \lambda + \frac{\|\mathbf{p}\|(\|\mathbf{p}\| - r)}{\|\mathbf{q}\|^2}$;

end Algorithm

We know that in general, Newton's method is not guaranteed to converge. If α satisfies that $\phi(\alpha)\phi''(\alpha) > 0$ then the iterations converge monotonically. Since $\phi''(\alpha)$ is positive in (λ_{\max}, ∞) , in fact $\phi(\alpha)$ is monotone decreasing in (λ_{\max}, ∞) , if $\phi(\alpha)$ is positive, then Newton's method is guaranteed to converge in (λ_{\max}, ∞) . If $\phi(\alpha)$ is negative, then next Newton iterate λ may less than λ_{\max} and converge to less than λ_{\max} . To make $-\mathbf{T} + \lambda \mathbf{I}$ positive definite, λ must stay in (λ_{\max}, ∞) . Hence, we need to find $\lambda > \lambda_{\max}$ such that $\phi(\lambda) > 0$. By using Gershgorin's theorem, we can find the upper bound of eigenvalues of \mathbf{T} . Let the initial λ be the upper bound of eigenvalue of \mathbf{T} , then $\lambda > \lambda_{\max}$. Repeat the following algorithm until we get $\lambda > \lambda_{\max}$ and $\phi(\lambda) \geq 0$.

Algorithm (Bisection)

$$\lambda = \lambda_{new} = \text{upper bound of the eigenvalues of } \mathbf{T}$$

```

evaluate  $\phi(\lambda)$ 
do while  $\lambda \leq \lambda_{\max}$  or  $\phi(\lambda) < 0$ 
    call Algorithm (Newton) to find  $\lambda_{new}$ 
    do while  $\lambda_{new} \leq \lambda_{\max}$ 
         $\lambda_{new} = \frac{\lambda_{new} + \lambda}{2}$ 
    end
     $\lambda = \lambda_{new}$  ( $\lambda > \lambda_{\max}$ )
    evaluate  $\phi(\lambda)$ 
end

```

end Algorithm

Now let us see how to check either $\lambda > \lambda_{\max}$ or $\lambda \leq \lambda_{\max}$. In general, let f_k denote the determinant of the k th leading submatrix of a matrix \mathbf{C} . Then

$$f_k = c_{kk}f_{k-1} - p_k f_{k-2}, \quad (4.17)$$

where $p_k = c_{kk-1}c_{k-1k}$ is the product between elements on opposite side of the diagonal and the initial values for the f_k are $f_0 = 1$ and $f_{-1} = 0$.

The recurrence (4.17) also provides a formula for the characteristic polynomial $\det(\mathbf{C} - \lambda \mathbf{I})$. If \mathbf{C} is a tridiagonal matrix, $\mathbf{C} - \lambda \mathbf{I}$ is the tridiagonal matrix obtained by subtracting λ from each diagonal element of \mathbf{C} . Letting $f_k(\lambda)$ denote the determinant of the k th leading submatrix for $\mathbf{C} - \lambda \mathbf{I}$, $f_k(\lambda)$ satisfies recurrence (4.17) except that c_{kk} is replaced by $c_{kk} - \lambda$:

$$f_k(\lambda) = (c_{kk} - \lambda)f_{k-1}(\lambda) - p_k f_{k-2}(\lambda), \quad (4.18)$$

where $f_0(\lambda) = 1$ and $f_{-1}(\lambda) = 0$.

We can check either $\lambda > \lambda_{\max}$ or $\lambda \leq \lambda_{\max}$ by the Sturm sequence property : For a $n \times n$ tridiagonal matrix with each p_k nonnegative, the number of eigenvalues smaller than λ is equal to the number of changes in sign for the sequence $f_0(\lambda), f_1(\lambda), \dots, f_n(\lambda)$. We know that \mathbf{T} is a symmetric tridiagonal matrix and so,

the p_k are always nonnegative : $p_k = t_{kk-1}t_{k-1k} = (t_{kk-1})^2$. Hence, the number of eigenvalues smaller than λ is equal to the number of sign changes. Thus, if there is no consecutive positive or negative signs, then λ is greater than λ_{\max} .

By above two algorithms (Newton and Bisection), we can find \mathbf{p} , the solution of (4.12) for $\mathbf{t} \neq \mathbf{0}$.

Let $\mathbf{y} = \mathbf{V}\mathbf{p}$, then this \mathbf{y} is the solution of (4.3) and of (4.1). Hence, we will take $\mathbf{x}_c + \mathbf{y}$ as the starting point.

Now, suppose $\mathbf{t} = \mathbf{0}$. Then

$$(4.12) \equiv \begin{array}{ll} \min & -\mathbf{z}^T \mathbf{T} \mathbf{z} \\ \text{subject to} & \mathbf{z}^T \mathbf{z} = r^2. \end{array} \quad (4.19)$$

The solution of (4.19), \mathbf{p} , is an eigenvector corresponding to the largest eigenvalue of \mathbf{T} and $\mathbf{V}\mathbf{p}$ is an eigenvector of the largest eigenvalue of \mathbf{PAP} . To find this eigenvector, we can use shifted power method. For this case, we want to make a biggest ball which may not in the convex set K , but the solution of (4.3) is in K . So, we multiple some scalar to the eigenvector of \mathbf{PAP} so that $\mathbf{x}_c + \mathbf{y} = \mathbf{x}_c + r \cdot \mathbf{V}\mathbf{p}$ is on the boundary of K . Let

$$r_i = \begin{cases} \frac{1-x_{c_i}}{\mathbf{V}\mathbf{p}_i} & \text{if } \mathbf{V}\mathbf{p}_i > 0 \\ \frac{-x_{c_i}}{\mathbf{V}\mathbf{p}_i} & \text{if } \mathbf{V}\mathbf{p}_i < 0 \end{cases}$$

and

$$r = \min_i(r_i).$$

Then this r will be the radius of the ball which may not in the convex set K , but the solution of (4.3) is in K . Hence, we will take $\mathbf{x}_c + \mathbf{y} = \mathbf{x}_c + r \cdot \mathbf{V}\mathbf{p}$ as the starting point.

1.2 The Solution of the Ball Centered at \mathbf{x}_c Containing K

$$\begin{aligned} r &= (\|x_c - \text{an extreme point of } K\|) \\ &= \sqrt{m - \frac{m^2}{n}} \end{aligned}$$

where x_c is a centroid of K , $\frac{m}{n}\mathbf{1}$.

By the same way, we can get the exact solution of the ball centered at \mathbf{x}_c containing K .

1.3 The Projection on K of the Solution of the Ball Containing K

If we get an exact solution of the ball which contains K , we need to evaluate the projection of the exact solution of the ball onto K . We will use the optimality conditions to evaluate the projection of the exact solution of the ball onto K .

$$\begin{aligned} & \min \frac{1}{2} \|\mathbf{x} - \mathbf{y}\|^2 \\ & \text{subject to } \mathbf{0} \leq \mathbf{x} \leq \mathbf{1}, \quad \mathbf{1}^T \mathbf{x} = m \end{aligned} \quad (4.20)$$

where $\mathbf{y} \in R^n$ is given. The solution to the problem (4.20) is $Proj_K(\mathbf{y})$. Since $\|\mathbf{x} - \mathbf{y}\|^2$ is strongly convex function, there exists a unique minimizer and the following first-order optimality conditions hold; there exist $\lambda \in R^1$ and $\mu, \nu \in R^n$ such that

$$\mathbf{x} - \mathbf{y} + \mu - \nu + \lambda \mathbf{1} = \mathbf{0}, \quad (4.21)$$

$$\mathbf{1}^T \mathbf{x} = m, \quad (4.22)$$

$$\mu^T (\mathbf{1} - \mathbf{x}) = 0, \quad \nu^T \mathbf{x} = 0 \quad (4.23)$$

$$\mu \geq \mathbf{0}, \quad \nu \geq \mathbf{0} \quad (4.24)$$

$$\mathbf{0} \leq \mathbf{x} \leq \mathbf{1} \quad (4.25)$$

Conversely, if \mathbf{x}, μ, ν and λ satisfies these conditions, then \mathbf{x} is the unique minimizer of (2.1).

For given $U, L \subset \{1, 2, \dots, n\}$, and $F = (U \cup L)^C$, we define

$$x_i = 1 \quad \text{if } i \in U, \quad x_i = 0 \quad \text{if } i \in L \quad (4.26)$$

$$\mu_i = 0, \quad \nu_i = 0 \quad \text{if } i \in F \quad (4.27)$$

$$\lambda = \frac{(|U| - m + \sum_{i \in F} y_i)}{|F|} \quad (4.28)$$

$$x_i = y_i - \lambda \quad \text{if } i \in F \quad (4.29)$$

$$\mu_i = y_i - 1 - \lambda \quad \text{if } i \in U, \quad \nu_i = \lambda - y_i \quad \text{if } i \in L \quad (4.30)$$

Then we can show that the conditions (4.21) – (4.23) are satisfied as before. Hence, we only need to show (4.24) and (4.25).

Let \mathbf{y} be the exact solution of the ball. Let $U_0 = L_0 = \emptyset$ and $F_0 = (U_0 \cup L_0)^C = \{1, 2, \dots, n\}$.

Define

$$\lambda_j = \frac{|U_j| - m + \sum_{i \in F_j} y_i}{|F_j|} \quad (4.31)$$

and for $i \in F_j$,

$$A_j = \{i \in F_j : a_i = y_i - 1 - \lambda_j \geq 0\} \quad (4.32)$$

$$B_j = \{i \in F_j : b_i = \lambda_j - y_i \geq 0\} \quad (4.33)$$

Let $\gamma_j = \sum_{i \in A_j} a_i$ and $\tau_j = \sum_{i \in B_j} b_i$. Define U_{j+1} and L_{j+1} as follow: For each j , if $\gamma_j \geq \tau_j$, then put each $i \in A_j$ into U_{j+1} , and if $\gamma_j < \tau_j$, then put each $i \in B_j$ into L_{j+1} . That is,

$$\begin{aligned} U_{j+1} &= U_j \cup A_j & \text{if } \gamma_j \geq \tau_j \\ U_{j+1} &= U_j & \text{otherwise,} \end{aligned} \quad (4.34)$$

$$\begin{aligned} L_{j+1} &= L_j \cup B_j & \text{if } \gamma_j < \tau_j \\ L_{j+1} &= L_j & \text{otherwise,} \end{aligned} \quad (4.35)$$

and

$$F_{j+1} = F_0 \setminus (U_{j+1} \cup L_{j+1}). \quad (4.36)$$

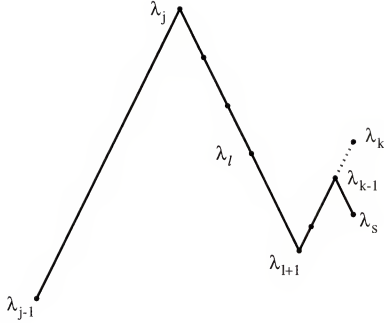
If we stop this iteration when $\gamma_s = \tau_s = 0$, i.e., $0 \leq x_i = y_i - \lambda_s \leq 1, \forall i \in F_s$, then by the following lemma, we can see that

$$\mu_i = y_i - 1 - \lambda_s \geq 0 \quad \forall i \in U_s,$$

$$\nu_i = \lambda_s - y_i \geq 0 \quad \forall i \in L_s$$

and

$$0 \leq \mathbf{x} \leq 1.$$

Figure 4.3: Convergence of λ

Lemma 18 *If $\gamma_j \geq \tau_j$, then $\lambda_j \geq \lambda_k$ for all $k > j$, and if $\gamma_j < \tau_j$, then $\lambda_j \leq \lambda_k$ for all $k > j$.*

Proof. First, let us consider the case $\gamma_j \geq \tau_j$. Let l be the first index greater than equal to j for which $\gamma_{l+1} < \tau_{l+1}$. Hence, $\gamma_i \geq \tau_i$ for all $j \leq i \leq l$. In (4.34), we take out A_i from F_i for each i . Thus, $\lambda_j > \dots > \lambda_l > \lambda_{l+1}$ and so, if $j < k \leq l+1$ then $\lambda_j > \lambda_k$. Let k be the first index greater than l for which $\gamma_k \geq \tau_k$. Hence, $\gamma_i < \tau_i$ for all $l+1 \leq i \leq k$. In (4.35), we remove B_i from F_i for each i . Thus, $\lambda_{l+1} < \dots < \lambda_{k-1} < \lambda_k$. We will show that $\lambda_j \geq \lambda_k$. Since $\lambda_j > \dots > \lambda_l$, it is enough to show that $\lambda_l \geq \lambda_k$. Note that $\lambda_j = \lambda_k$ holds if $j = l$ and $\gamma_j = \tau_j$ and $A_i = B_i = \emptyset$ for $j < i < k$. By way of contradiction, suppose $\lambda_k > \lambda_l$. Without loss of generality, we can assume $\lambda_k > \lambda_l \geq \lambda_{k-1}$, i.e.,

$$\lambda_{l+1} < \dots < \lambda_{k-1} \leq \lambda_l < \lambda_k. \quad (4.37)$$

Then

$$\lambda_k = \frac{|U_k| - m + \sum_{i \in F_k} y_i}{|F_k|} \quad (4.38)$$

$$= \frac{|U_l| - m + \sum_{i \in F_l} y_i - \sum_{i \in A_l} (y_i - 1) - \sum_{q=l+1}^{k-1} \sum_{i \in B_q} y_i}{|F_k|} \quad (4.39)$$

$$= \frac{|F_l| \lambda_l - \sum_{i \in A_l} (y_i - 1) - \sum_{q=l+1}^{k-1} \sum_{i \in B_q} y_i}{|F_k|}. \quad (4.40)$$

Hence,

$$\lambda_k - \lambda_l = \frac{(|F_l| - |F_k|) \lambda_l - \sum_{i \in A_l} (y_i - 1) - \sum_{q=l+1}^{k-1} \sum_{i \in B_q} y_i}{|F_k|}. \quad (4.41)$$

Since $\lambda_k > \lambda_l$,

$$(|F_l| - |F_k|) \lambda_l - \sum_{i \in A_l} (y_i - 1) - \sum_{q=l+1}^{k-1} \sum_{i \in B_q} y_i > 0 \quad (4.42)$$

On the other hand, since $\gamma_l \geq \tau_l$, we have

$$\sum_{i \in A_l} (y_i - 1 - \lambda_l) \geq \sum_{i \in B_l} (\lambda_l - y_i). \quad (4.43)$$

Hence by (4.42), we have

$$\begin{aligned} & (|F_l| - |F_k|) \lambda_l - \sum_{i \in A_l} (y_i - 1) - \sum_{q=l+1}^{k-1} \sum_{i \in B_q} y_i > 0 \\ \Rightarrow & - \sum_{i \in A_l} (y_i - 1 - \lambda_l) - \sum_{q=l+1}^{k-1} \sum_{i \in B_q} (y_i - \lambda_q) > 0 \\ \Rightarrow & \sum_{q=l+1}^{k-1} \sum_{i \in B_q} (\lambda_q - y_i) - \sum_{i \in A_l} (y_i - 1 - \lambda_l) > 0 \\ \Rightarrow & \sum_{q=l+1}^{k-1} \sum_{i \in B_q} (\lambda_q - y_i) - \sum_{i \in B_l} (\lambda_l - y_i) > 0 \end{aligned} \quad (4.44)$$

Since for $l+1 \leq q \leq k-1$, F_q is monotone decreasing and $\lambda_q \leq \lambda_j$, we have

$$B_q \subset B_l$$

and so,

$$\sum_{q=l+1}^{k-1} \sum_{i \in B_q} (\lambda_q - y_i) \leq \sum_{q=l+1}^{k-1} \sum_{i \in B_q} (\lambda_l - y_i) < \sum_{i \in B_l} (\lambda_l - y_i). \quad (4.45)$$

This contradicts (4.44). Therefore, $\lambda_l \geq \lambda_k$ if k satisfies $\gamma_i < \tau_i$ for all $l+1 \leq i \leq k$.

Hence $\lambda_j \geq \lambda_k$.

Now consider a general $k > j$. We will use mathematical induction to show that $\lambda_j \geq \lambda_k$ for all $k > j$. Let k_i be the i -th number after j such that $\gamma_{k_i-1} < \tau_{k_i-1}$ and $\gamma_{k_i} \geq \tau_{k_i}$. We have seen that $\lambda_j \geq \lambda_k$ for $j < k \leq k_1$. Suppose it is true for $j < k \leq k_i$. Then $\lambda_j \geq \lambda_{k_i}$. Let $k_i < k \leq k_{i+1}$, then by the above argument, we have $\lambda_{k_i} \geq \lambda_k$. Hence, $\lambda_j \geq \lambda_k$ for all $k > j$. Similarly, we can show that if $\gamma_j < \tau_j$, then $\lambda_j \leq \lambda_k$ for all $k > j$. \square

Hence, by Lemma 18,

$$\mu_i = y_i - 1 - \lambda_k \geq 0 \quad \forall i \in U_j, \forall k \geq j \quad (4.46)$$

$$\nu_i = \lambda_k - y_i \geq 0 \quad \forall i \in L_j, \forall k \geq j. \quad (4.47)$$

Let us stop this process if $\gamma_j = \tau_j = 0$. Then since F_0 is finite, this must be terminated in finite steps. Let s is the first number such that $\gamma_s = \tau_s = 0$, then define $U = U_s = U_{s+1}$, $L = L_s = L_{s+1}$, $F = F_s = F_0 \setminus (U_s \cup L_s)$ and $\lambda = \lambda_s$. Hence, for all i ,

$$\mu_i \geq 0 \text{ and } \nu_i \geq 0$$

and

$$0 \leq x_i \leq 1.$$

Thus, \mathbf{x} in K such that

$$x_i = 1 \text{ if } i \in U, \quad x_i = 0 \text{ if } i \in L \quad \text{and} \quad x_i = y_i - \lambda \text{ if } i \in F$$

is the projection of \mathbf{y} onto K .

2 Starting Procedure

(1) First of all, put all indices in F and evaluate the new direction vector g with x_0 . Then $U = \emptyset$ and $F = \emptyset$.

(2) Evaluate $\mathbf{g} = -\nabla f(\mathbf{x}_0) = -(\mathbf{A}\mathbf{1} - 2\mathbf{A}\mathbf{x}_0)$.

(3) By using Index Decision procedure, determine U , L and F at $\alpha = \alpha_0^+$ with $\mathcal{U} = \{i : x_0, i = 1\}$, $\mathcal{L} = \{i : x_0, i = 0\}$.

(4) Evaluate $f'(\mathbf{x}(0))$ and $f''(\mathbf{x}(0))$

$$f'(\mathbf{x}(\alpha))|_{\alpha=0} = \mathbf{g}_F^T \mathbf{A}_F [(1+2\bar{c})\mathbf{1}_F - 2\mathbf{x}_{0_F} + (\mathbf{1}_L - \mathbf{1}_U)] \\ - \mathbf{1}_F^T \mathbf{A}_F \bar{g} [(1+2\bar{c})\mathbf{1}_F - 2\mathbf{x}_{0_F} + (\mathbf{1}_L - \mathbf{1}_U)]$$

$$f''(\mathbf{x}(\alpha))|_{\alpha=0} = -2(\mathbf{g}_F^T \mathbf{A}_{FF} - \bar{g} \mathbf{1}_F^T \mathbf{A}_{FF})(\mathbf{g}_F - \bar{g} \mathbf{1}_F)$$

(5) If $g_i \geq \bar{g} \forall i \in U_0$, $g_i \leq \bar{g} \forall i \in L_0$ and $g_i = \bar{g} \forall i \in F_0$, then go to error evaluation, else go to continuing procedure.

3 Continuing Procedure

(1). By definition of α_{j+1} , find new $\alpha(\alpha_{j+1})$ which is larger than old $\alpha(\alpha_j)$.

$$\alpha_{j+1} = \min \left\{ \begin{array}{l} \sup\{\alpha : 0 < x_i(F_j, \alpha) < 1 \forall i \in F_j, x_i(F_j, \alpha_j^+) \neq 0 \text{ or } 1\}, \\ \sup\{\alpha : \mu_i(\alpha) > 0 \forall i \in U_j, \mu_i(F_j, \alpha_j^+) \neq 0\}, \\ \sup\{\alpha : \nu_i(\alpha) > 0 \forall i \in L_j, \nu_i(F_j, \alpha_j^+) \neq 0\} \end{array} \right\}$$

(2). By using the following procedure, Index Decision, determine new U , L and F at $\alpha = \alpha_{j+1}$.

3.1 Index Decision

Define \mathcal{U} and \mathcal{L} such that

$$\text{if } \mu_i(\alpha_{j+1}) = 0, x_i(\alpha_{j+1}) = 1 \quad \text{then } i \in \mathcal{U},$$

$$\text{if } \nu_i(\alpha_{j+1}) = 0, x_i(\alpha_{j+1}) = 0 \quad \text{then } i \in \mathcal{L}.$$

To determine new U , L and F at $\alpha = \alpha_{j+1}$ repeat the following until $\gamma = \tau = 0$. Let

$$F = F \cup \mathcal{U} \cup \mathcal{L}, \quad U = U \setminus \mathcal{U}, \quad L = L \setminus \mathcal{L},$$

$$\gamma = \sum_{i \in \mathcal{U}, g_i > \bar{g}} (g_i - \bar{g}), \quad \tau = \sum_{i \in \mathcal{L}, g_i < \bar{g}} (\bar{g} - g_i) \quad \text{where } \bar{g} = \frac{\sum_{i \in F} g_i}{|F|}, .$$

If $\gamma \geq \tau$,

$$U = U \cup \{i \in \mathcal{U} : g_i > \bar{g}\}$$

$$L^{l+1} = L^l$$

$$\mathcal{U} = \mathcal{U} \setminus \{i \in \mathcal{U} : g_i > \bar{g}\}$$

If $\gamma < \tau$,

$$\begin{aligned} L &= L \cup \{i \in \mathcal{L} : g_i < \bar{g}\} \\ L^{l+1} &= L^l \\ \mathcal{L} &= \mathcal{L} \setminus \{i \in \mathcal{L} : g_i < \bar{g}\} \\ F &= F \setminus (\{i \in \mathcal{U} : g_i > \bar{g}\} \cup \{i \in \mathcal{L} : g_i < \bar{g}\}). \end{aligned}$$

Let us stop this process if $\gamma = \tau = 0$. Then

$$F = F \cup \mathcal{U} \cup \mathcal{L}, \quad U = U \text{ and } L = L$$

□

(3) Since \mathbf{x} is piecewise linear with respect to α , evaluate $f'(\mathbf{x}(\alpha_{j+1}^+))$, $f'(\mathbf{x}(\alpha_{j+1}^-))$, $f'(\mathbf{x}(0)) = f'(\mathbf{x}(\alpha))|_{\alpha=0}$ and $f''(\mathbf{x}(\alpha_{j+1}))$.

(4) If we reach the extreme point of g on K or some $\alpha = \alpha_j$ such that

$$g_i \geq \bar{g} \quad \forall i \in U_j, \quad g_i \leq \bar{g} \quad \forall i \in L_j \text{ and } g_i = \bar{g} \quad \forall i \in F_j,$$

then stop this continuing procedure and go to starting procedure with the global minimizer $x(\alpha)$ for given g as new starting point x_0 .

If old derivative of f , $f'(x(\alpha_{j-1}))$, is less than 0 and derivative of f at α_j^- , $f'(x(\alpha_j^-))$, is greater than 0, then local minimum of f occurs at $x(\alpha)$ where $\alpha = \frac{f'(x(0))}{f''(x(\alpha))}$.

If $f'(x(\alpha_j^-)) \leq 0$ and $f'(x(\alpha_j)) > 0$, or $f'(x(\alpha_j^-)) \leq 0$ and $f'(x(\alpha_j)) = 0$, then local minimum of $f(x(\alpha))$ occurs at $\alpha = \alpha_j$.

If we can find another local minimum of $f(x(\alpha))$ in the interval $[\alpha_j, \alpha_{j+1}]$, compare this with the best value of $f(x(\alpha))$ in the interval $[0, \alpha_j]$, and then keep the smaller value $f(x(\alpha))$ as the global minimum and the corresponding $\mathbf{x}(\alpha)$ as the global minimizer for a given \mathbf{g} . Then repeat this continuing procedure until meet stopping point or extreme point of g on K .

4 Error Evaluation

In this error evaluation procedure, we will check the first order condition and define $\tilde{\mu}$ and $\tilde{\lambda}$.

Given a scalar $\tilde{\lambda}$, we define the vector

$$\tilde{\mu}(\mathbf{x}, \tilde{\lambda}) = \mathbf{A}\mathbf{1} - 2\mathbf{A}\mathbf{x} + \tilde{\lambda}\mathbf{1}.$$

To satisfy the first order condition, $\tilde{\mu}_i = 0$ if $i \in F$, $\tilde{\mu}_i < 0$ if $i \in U$ and $\tilde{\mu}_i > 0$ if $i \in L$. Hence,

$$\begin{aligned} \tilde{\lambda} &= -\frac{\mathbf{1}_F^T (\mathbf{A}\mathbf{1} - 2\mathbf{A}\mathbf{x})_F}{|F|} \quad \text{if } |F| \neq 0, \\ \tilde{\lambda} &= -\frac{\max(\mathbf{A}\mathbf{1} - 2\mathbf{A}\mathbf{x})_U + \min(\mathbf{A}\mathbf{1} - 2\mathbf{A}\mathbf{x})_L}{2} \quad \text{if } |F| = 0. \end{aligned}$$

Now evaluate the error.

$$\begin{aligned} \text{Error} &= \text{error in } F \text{ component} + \text{error in } U \text{ component} + \text{error in } L \text{ component} \\ &= \|F \text{ component}\|_2 + \|U \text{ component}\|_2 + \|L \text{ component}\|_2 \\ &= \|\tilde{\mu}_{F+}\| + \|\tilde{\mu}_{U+}\| + \|\tilde{\mu}_{L+}\| \end{aligned}$$

where $\tilde{\mu}_{F+} = \{i \in F : \tilde{\mu}_i \neq 0\}$, $\tilde{\mu}_{U+} = \{i \in U : \tilde{\mu}_i > 0\}$ and $\tilde{\mu}_{L+} = \{i \in L : \tilde{\mu}_i < 0\}$.

If error = 0, then it satisfies the first order condition and so go to perturbation procedure to check the second order conditions and so on. Otherwise, it is not a local minimizer, and so in the perturbation procedure we do not have to check the second order conditions.

5 Perturbation Procedure

In this perturbation procedure, define *Upper* and *Lower* are active sets and *Free* is the set of free indices, i.e., $Upper = \{i : x_i = 1\}$, $Lower = \{i : x_i = 0\}$ and $Free = \{i : 0 < x_i < 1\}$ which are different from *U*, *L* and *F*, and check the following second order conditions :

- (1) For each i and $j \in Free$, $a_{ij} = 1$.
- (2) Consider the three sets $Upper_0$, $Lower_0$, and $Free$, where $Upper_0 = \{i : x_i = 1 \text{ and } \mu_i = 0\}$ and $Lower_0 = \{i : x_i = 0 \text{ and } \mu_i = 0\}$. For each i and j in two different sets, we have $a_{ij} = 1$.

If it satisfies above second order conditions, then it is a local minimizer, if does not, it is a stationary point. If it is a local minimizer with no free index, *i.e.*, all components of local minimizer are 0 or 1, then go to the Kernighan-Lin method for local minimizer. If it is a stationary point, then perturb the components which violate the condition. If it is a local minimizer but has free indices, then perturb free components by using the power method.

Perturb the point in order to at least one more component become boundary. Then go to starting procedure with the perturbed point.

5.1 Perturbation by Using Power Method

We want to perturb the free components of \mathbf{x} so that at least one free component of \mathbf{x} become 0 or 1, by adding a vector \mathbf{z} such that $\max_{\mathbf{z}} \mathbf{z}^T \mathbf{A}_{FF} \mathbf{z}$ and $\mathbf{1}^T \mathbf{z} = 0$ where \mathbf{A}_{FF} is the matrix of free components. Let $\mathbf{P} = \mathbf{I} - \frac{1}{k} \mathbf{1} \mathbf{1}^T$ be the projection to $\mathbf{1}$ where k is the size of \mathbf{A}_{FF} . Then we notice that $\mathbf{P}^2 = \mathbf{P}$ and $\mathbf{P}^T = \mathbf{P}$.

$$\mathbf{P}^2 = \left(\mathbf{I} - \frac{1}{k} \mathbf{1} \mathbf{1}^T \right) \left(\mathbf{I} - \frac{1}{k} \mathbf{1} \mathbf{1}^T \right) = \mathbf{I} - \frac{2}{k} \mathbf{1} \mathbf{1}^T + \frac{1}{k^2} \mathbf{1} \mathbf{1}^T \mathbf{1} \mathbf{1}^T = \mathbf{I} - \frac{1}{k} \mathbf{1} \mathbf{1}^T = \mathbf{P}.$$

$$\mathbf{P}^T = \left(\mathbf{I} - \frac{1}{k} \mathbf{1} \mathbf{1}^T \right)^T = \left(\mathbf{I} - \frac{1}{k} \mathbf{1} \mathbf{1}^T \right) = \mathbf{P}.$$

Lemma 19 *Let A be a symmetric $k \times k$ matrix and let $\lambda_1 \leq \dots \leq \lambda_k$ be its (real) eigenvalues. Then*

$$\lambda_1 \|\mathbf{y}\|^2 \leq \mathbf{y}^T A \mathbf{y} \leq \lambda_n \|\mathbf{y}\|^2 \text{ for all } \mathbf{y} \in \mathbb{R}^k.$$

Proof. See p545 in [3]. □

Theorem 20 *If $\text{diag}(\mathbf{A}) = \mathbf{I}$, $a_{ij} = 0$ or 1 for all i, j and $a_{ij} = 1$ for some $i \neq j$, then*

$$\min_{\|\mathbf{x}\|=1} \mathbf{x}^T \mathbf{P} \mathbf{A} \mathbf{P} \mathbf{x} = \min_{\|\mathbf{x}\|=1, \mathbf{x} \perp \mathbf{1}} \mathbf{x}^T \mathbf{A} \mathbf{x}, \quad \mathbf{P} = \mathbf{I} - \frac{1}{k} \mathbf{1} \mathbf{1}^T.$$

Proof. Let $\|\cdot\|$ be the Euclidean norm and $\|\mathbf{x}\| = 1$. Define \mathbf{x}_P is the projection of \mathbf{x} on $\mathbf{1}$ and $\mathbf{x}_N = (\mathbf{x} - \mathbf{x}_P)$ is perpendicular to $\mathbf{1}$. Then $\mathbf{P}\mathbf{x}_P = \mathbf{0}$ and $\mathbf{P}\mathbf{x}_N = \mathbf{x}_N$.

$$\min_{\|\mathbf{x}\|=1} \mathbf{x}^T \mathbf{P} \mathbf{A} \mathbf{P} \mathbf{x} = \min_{\|\mathbf{x}_N\|^2 + \|\mathbf{x}_P\|^2 = 1} \mathbf{x}_N^T \mathbf{A}_{FF} \mathbf{x}_N = \min_{\|\mathbf{x}_N\| \leq 1, \mathbf{x} \perp \mathbf{1}} \mathbf{x}_N^T \mathbf{A} \mathbf{x}_N.$$

Under the hypothesis, all entries of A are one or zero with $a_{ij} = 1$ for some $i \neq j$ and $\mathbf{1}^T \mathbf{x} = 0$. Hence, we have $\mathbf{x}^T \mathbf{A} \mathbf{x} = 0$ with $x_i = 1$, $x_j = -1$, and $x_l = 0$ for $l \neq i, j$. By lemma 19, A has eigenvalue(s) which is(are) less than or equal to zero. We claim that

$$\min_{\|\mathbf{x}_N\| \leq 1, \mathbf{x} \perp \mathbf{1}} \mathbf{x}_N^T \mathbf{A} \mathbf{x}_N = \min_{\|\mathbf{x}_N\|=1, \mathbf{x}_N \perp \mathbf{1}} \mathbf{x}_N^T \mathbf{A} \mathbf{x}_N \quad (4.48)$$

$\min_{\|\mathbf{x}_N\| \leq 1, \mathbf{x}_N \perp \mathbf{1}} \mathbf{x}_N^T \mathbf{A} \mathbf{x}_N \leq \min_{\|\mathbf{x}_N\|=1, \mathbf{x}_N \perp \mathbf{1}} \mathbf{x}_N^T \mathbf{A} \mathbf{x}_N$ is obvious. Now we want to show the other inequality. If the norm of the solution of left hand side in (4.48) is one, we have equality. Suppose \mathbf{x}_N is the solution of left hand side in (4.48) with $\|\mathbf{x}_N\| < 1$, then there exist $\beta > 1$ such that $\|\beta \mathbf{x}_N\| = 1$. Let $\mathbf{y} = \beta \mathbf{x}_N$. Since $\beta > 1$ and $\mathbf{x}_N^T \mathbf{A} \mathbf{x}_N < 0$,

$$\mathbf{y}^T \mathbf{A} \mathbf{y} = \beta^2 \mathbf{x}_N^T \mathbf{A} \mathbf{x}_N \leq \mathbf{x}_N^T \mathbf{A} \mathbf{x}_N.$$

Thus, we hold (4.48). Since, $\mathbf{P}\mathbf{x}_P = \mathbf{0}$ and $\mathbf{P}\mathbf{x}_N = \mathbf{x}_N$, we have

$$\min_{\|\mathbf{x}\|=1} \mathbf{x}^T \mathbf{P} \mathbf{A} \mathbf{P} \mathbf{x} = \min_{\|\mathbf{x}\|=1, \mathbf{x} \perp \mathbf{1}} \mathbf{x}^T \mathbf{A} \mathbf{x}, \quad \mathbf{P} = \mathbf{I} - \frac{1}{k} \mathbf{1} \mathbf{1}^T.$$

□

Corollary 21

$$\min_{\|\mathbf{x}_N\|=1, \mathbf{x}_N \perp \mathbf{1}} \mathbf{x}_N^T \mathbf{A} \mathbf{x}_N \geq \min_{\|\mathbf{x}\|=1} \mathbf{x}^T \mathbf{A} \mathbf{x}.$$

Proof. It is obvious because left hand side has less conditions. □

Hence by theorem 20 and corollary 21, the smallest eigenvalue of $\mathbf{P} \mathbf{A}_{FF} \mathbf{P}$ can be bounded from below by the smallest eigenvalue of \mathbf{A}_{FF} which can be estimated

by Gershgorin's theorem.

$$|\lambda - 1| \leq \sum_{j=1, j \neq i}^k |a_{ij}| \Leftrightarrow 1 - \sum_{j=1, j \neq i}^k |a_{ij}| \leq \lambda. \quad (4.49)$$

Let λ_{\min} and λ_{\max} be the most negative and positive eigenvalues of $\mathbf{PA}_{FF}\mathbf{P}$ respectively.

Theorem 22 *If $\alpha \geq -\frac{\min_i(1 - \sum_{j=1, j \neq i}^k |a_{ij}|)}{2}$, then the dominant eigenvalue of $\mathbf{M} = \mathbf{PA}_{FF}\mathbf{P} + \alpha\mathbf{I}$ is the most positive eigenvalue of \mathbf{M} .*

Proof. We know by (4.49) that all eigenvalues of $\mathbf{PA}_{FF}\mathbf{P}$ are bounded from below by $1 - \sum_{j=1, j \neq i}^k |a_{ij}|$. By Lemma 19, we know that $\lambda_{\min} \leq 0 \leq \lambda_{\max}$. Let λ_m and λ_M be the most negative and positive eigenvalues of \mathbf{M} respectively, then $\lambda_m = \lambda_{\min} + \alpha$ and $\lambda_M = \lambda_{\max} + \alpha$. Hence,

$$\begin{aligned} \alpha \geq -\frac{\min_i(1 - \sum_{j=1, j \neq i}^k |a_{ij}|)}{2} &\Rightarrow \alpha \geq -\frac{\lambda_{\min}}{2} \\ &\Rightarrow \alpha \geq -\frac{\lambda_{\min} + \lambda_{\max}}{2} \\ &\Rightarrow -(\lambda_{\min} + \alpha) \leq \lambda_{\max} + \alpha \\ &\Rightarrow -\lambda_m \leq \lambda_M \\ &\Rightarrow |\lambda_m| \leq |\lambda_M| \end{aligned}$$

Hence if we choose α such that $\alpha \geq -\frac{\lambda_{\min}}{2}$, then the dominant eigenvalue of \mathbf{M} is the most positive eigenvalue of \mathbf{M} . \square

Now we apply a small number of iterations of the power method as follows:

$$\begin{aligned} \mathbf{z}_{k+1} &= \mathbf{M}\mathbf{z}_k, \quad \mathbf{M} = \mathbf{PA}_{FF}\mathbf{P} + \alpha\mathbf{I} \\ \mathbf{u} &= \frac{\mathbf{z}_{k+1}}{\|\mathbf{z}_{k+1}\|} \\ \mathbf{z}_{k+1} &= \mathbf{u} \end{aligned}$$

Let us stop this power method if $\|\mathbf{z}_{k+1} - \mathbf{z}_k\| \approx 0$. Then \mathbf{u} is the eigenvector corresponding to the dominant eigenvalue of \mathbf{M} . Since eigenvectors for \mathbf{M} are equal

to the eigenvectors of $\mathbf{P}\mathbf{A}_{FF}\mathbf{P}$, \mathbf{u} is the eigenvector corresponding to the dominant eigenvalue of \mathbf{A}_{FF} .

If \mathbf{x} whose components are all 0 or 1, does not satisfy the conditions of a local minimizer or $0 < x_i < 1$ for some i and so \mathbf{x} is perturbed, then go to the starting procedure with the perturbed \mathbf{x} . If \mathbf{x} is a local minimizer and $x_i = 0$ or 1 for all i , then go to the exchange method for a local minimizer.

6 Exchange Method for a Local Minimizer

Now, we reach a local minimizer whose components are all 0 or 1. In this section, thus we assume that all components of the local minimizer \mathbf{x} are 0 or 1.

6.1 Kernighan-Lin Method

Let us see the Kernighan-Lin method (see [13]). Suppose that the matrix \mathbf{C} is symmetric and diagonal is zero. Let $H = \{i : x_i = 0\}$ and $B = \{i : x_i = 1\}$. Define for each $h_1 \in H$, an *external cost* E_{h_1} by

$$E_{h_1} = \sum_{b \in B} c_{bh_1}$$

and an *internal cost* I_{h_1} by

$$I_{h_1} = \sum_{h \in H} c_{hh_1}.$$

Similarly, define E_{b_1} , I_{b_1} for each $b_1 \in B$. Let $D_z = E_z - I_z$ for all $z \in S = H \cup B$; D_z is the difference between external and internal costs.

Lemma 23 (Kernighan-Lin method) *Consider any $h \in H$, $b \in B$. If $b_1 \in B$ and $h_1 \in H$ are exchanged, the gain (that is, the reduction in cost) is precisely*

$$g_{b_1 h_1} = D_{b_1} + D_{h_1} - 2c_{b_1 h_1} \quad (4.50)$$

Furthermore, we can recalculate the D value by

$$\begin{aligned} D'_b &= D_b - 2c_{bh_1} + 2c_{bb_1} \quad \text{for } b \in B - \{b_1\} \\ D'_h &= D_h + 2c_{hh_1} - 2c_{hb_1} \quad \text{for } h \in H - \{h_1\}. \end{aligned} \quad (4.51)$$

□

6.2 Improved Kernighan-Lin Method

In the Kernighan-Lin method, if the maximum gain is less than equal to zero, then stop the process, but we want to continue the process even if the maximum gain is zero. Since even if the maximum gain is zero in current iteration, we may have positive gain in the next process with the exchange of components whose gain is zero in current iteration. So we concern all $b \in B$ and $h \in H$ such that $g_{bh} \geq 0$. We will compare the numerical result of the Kernighan and improved Kernighan-Lin methods for the quadratic problem $f(\mathbf{x}) = (1 - \mathbf{x})^T \mathbf{A} \mathbf{x}$. Hence the formulas of $D'(b)$, $D'(h)$ and $g_{b_1 h_1}$ for improved will be exactly same as those for the Kernighan-Lin method. But the updated sets, $newH = H + \{b_1\} - \{h_1\}$ and $newB = B + \{h_1\} - \{b_1\}$ for the improved Kernighan-Lin method are different from the sets $newH = H - \{h_1\}$ and $newB = B - \{b_1\}$ for the Kernighan-Lin method. Since we continue the process even if the maximum gain is zero, we need to choose which components whose gain is zero in the current iteration will be exchanged. Since $newH \cup newB$ is equal to $H \cup B$ in the improved Kernighan-Lin method, we also need to control the infinite loop. We will see the details later in the section, exchange method procedure.

6.3 Formulas for Exchange Method

The formulas in this section (4.52) and (4.54) are the formulas for the improved Kernighan-Lin method for the quadratic problem $f(\mathbf{x}) = (1 - \mathbf{x})^T \mathbf{A} \mathbf{x}$.

Lemma 24 *Let \mathbf{x} be an n -vector with all components 0 or 1 and $f(\mathbf{x}) = (1 - \mathbf{x})^T \mathbf{A} \mathbf{x}$. If \mathbf{y} be an n -vector such that $y_l = x_k = 1$, $y_k = x_l = 0$ and $y_i = x_i \forall i \neq l, k$, then the gain $g = f(\mathbf{x}) - f(\mathbf{y})$ is*

$$g = \left(\sum_{i=1}^n a_{ki} - 2 \sum_{i=1}^n a_{ki} x_i \right) + \left(2 \sum_{i=1}^n a_{li} x_i - \sum_{i=1}^n a_{li} \right) + 2 - 2a_{kl}. \quad (4.52)$$

Proof.

$$(1 - \mathbf{x})^T \mathbf{A} \mathbf{x}$$

$$\begin{aligned}
&= \sum_{i=1}^n (1 - x_i) \left(\sum_{j=1}^n a_{ij} x_j \right) \\
&= \sum_{i=1}^n (1 - x_i) \left[\left(\sum_{j \neq l, k} a_{ij} x_j \right) + a_{il} x_l + a_{ik} x_k \right] \\
&= \sum_{i \neq l, k} (1 - x_i) \left[\left(\sum_{j \neq l, k} a_{ij} x_j \right) + a_{il} x_l + a_{ik} x_k \right] \\
&\quad + (1 - x_l) \left[\left(\sum_{j \neq l, k} a_{lj} x_j \right) + a_{ll} x_l + a_{lk} x_k \right] \\
&\quad + (1 - x_k) \left[\left(\sum_{j \neq l, k} a_{kj} x_j \right) + a_{kl} x_l + a_{kk} x_k \right] \\
&= \sum_{i \neq l, k} (1 - x_i) \left(\sum_{j \neq l, k} a_{ij} x_j \right) + \sum_{i \neq l, k} (1 - x_i) (a_{il} x_l + a_{ik} x_k) \\
&\quad + (1 - x_l) \left[\left(\sum_{j \neq l, k} a_{lj} x_j \right) + a_{ll} x_l + a_{lk} x_k \right] \\
&\quad + (1 - x_k) \left[\left(\sum_{j \neq l, k} a_{kj} x_j \right) + a_{kl} x_l + a_{kk} x_k \right]
\end{aligned}$$

Since $y_l = x_k$, $y_k = x_l$ and $y_i = x_i \ \forall i \neq l, k$,

$$\begin{aligned}
&(\mathbf{1} - \mathbf{y})^T \mathbf{A} \mathbf{y} \\
&= \sum_{i \neq l, k} (1 - x_i) \left(\sum_{j \neq l, k} a_{ij} x_j \right) + \sum_{i \neq l, k} (1 - x_i) (a_{il} x_k + a_{ik} x_l) \\
&\quad + (1 - x_k) \left[\left(\sum_{j \neq l, k} a_{lj} x_j \right) + a_{ll} x_k + a_{lk} x_l \right] \\
&\quad + (1 - x_l) \left[\left(\sum_{j \neq l, k} a_{kj} x_j \right) + a_{kl} x_k + a_{kk} x_l \right]
\end{aligned}$$

Hence, the gain, old cost - new cost = $f(\mathbf{x}) - f(\mathbf{y})$, is

$$\begin{aligned}
g &= (\mathbf{1} - \mathbf{x})^T \mathbf{A} \mathbf{x} - (\mathbf{1} - \mathbf{y})^T \mathbf{A} \mathbf{y} \\
&= \sum_{i \neq l, k} (1 - x_i) [a_{il}(x_l - x_k) + a_{ik}(x_k - x_l)] \\
&\quad + (1 - x_l) \left[\left(\sum_{j \neq l, k} a_{lj} x_j \right) - \left(\sum_{j \neq l, k} a_{kj} x_j \right) + (a_{ll} - a_{kk})x_l + (a_{lk} - a_{kl})x_k \right]
\end{aligned}$$

$$\begin{aligned}
& + (1 - x_k) \left[\left(\sum_{j \neq l, k} a_{kj} x_j \right) - \left(\sum_{j \neq l, k} a_{lj} x_j \right) + (a_{kl} - a_{lk}) x_l + (a_{kk} - a_{ll}) x_k \right] \\
& = \sum_{i \neq l, k} (1 - x_i) (a_{ik} - a_{il}) + \sum_{j \neq l, k} a_{lj} x_j - \sum_{j \neq l, k} a_{kj} x_j \\
& \quad (\text{since } x_l = 0, x_k = 1, a_{lk} = a_{kl}) \\
& = \sum_{i \neq l, k} (1 - 2x_i) (a_{ik} - a_{il}) \\
& = \sum_{i=1}^n (1 - 2x_i) (a_{ik} - a_{il}) - (1 - 2x_l) (a_{lk} - a_{ll}) - (1 - 2x_k) (a_{kk} - a_{kl}) \\
& = \sum_{i=1}^n (1 - 2x_i) (a_{ik} - a_{il}) + 2 - 2a_{kl} \quad \text{since } x_l = 0, x_k = 1, a_{ll} = a_{kk} = 1 \\
& = \left(\sum_{i=1}^n a_{ki} - 2 \sum_{i=1}^n a_{ki} x_i \right) + \left(2 \sum_{i=1}^n a_{li} x_i - \sum_{i=1}^n a_{li} \right) + 2 - 2a_{kl}
\end{aligned}$$

□

Let us define $D(i) = \sum_{j=1}^n a_{ij} - 2 \sum_{j=1}^n a_{ij} x_j$ for all i , and $H = \{i : x_i = 0\}$, $B = \{i : x_i = 1\}$. Let us define g_{bh} for $h \in H, b \in B$, as the gain by exchanging $b \in B$ and $h \in H$, that is,

$$g_{bh} = \left(\sum_{i=1}^n a_{ki} - 2 \sum_{i=1}^n a_{ki} x_i \right) + \left(2 \sum_{i=1}^n a_{li} x_i - \sum_{i=1}^n a_{li} \right) + 2 - 2a_{kl}.$$

Lemma 25 *If $b_1 \in B$ and $h_1 \in H$ are exchanged i.e. new $H = H' = H + \{b_1\} - \{h_1\}$ and new $B = B' = B + \{h_1\} - \{b_1\}$, then for $i \in B'$ and $i \in H'$,*

$$g = D(b_1) - D(h_1) + 2 - 2a_{b_1 h_1} \quad (4.53)$$

and $D(i)$ can be recalculated by

$$D(i)' = D(i) + 2a_{ib_1} - 2a_{ih_1}. \quad (4.54)$$

Proof. (4.53) is followed by (4.52) in Lemma 24. Since $b_1 \in B$ and $h_1 \in H$ are exchanged, $x_{b_1} = 1$ and $x_{h_1} = 0$. Hence,

$$D(i)' = \left(\sum_{j=1}^n a_{ij} - 2 \sum_{j=1}^n a_{ij} x_j \right)'$$

$$\begin{aligned}
&= \sum_{j=1}^n a_{ij} - 2 \left(\sum_{j=1}^n a_{ij}x_j - a_{ib_1}x_{b_1} + a_{ib_1}x_{h_1} - a_{ih_1}x_{h_1} + a_{ih_1}x_{b_1} \right) \\
&= \sum_{j=1}^n a_{ij} - 2 \sum_{j=1}^n a_{ij}x_j + 2a_{ib_1} - 2a_{ih_1} \\
&= D(i) + 2a_{ib_1} - 2a_{ih_1}
\end{aligned}$$

□

We want to show that our formulas for exchange method are equivalent to the formulas in the improved Kernighan-Lin method for the quadratic programming problem $f(\mathbf{x}) = (1 - \mathbf{x})^T \mathbf{A} \mathbf{x}$. Since the matrix \mathbf{C} in the Kernighan-Lin method is symmetric and all diagonals are zero, the relation between \mathbf{C} and our matrix \mathbf{A} is $\mathbf{A} = \mathbf{C} + \mathbf{I}$.

Lemma 26 *For given conditions,*

$$D_h = -D(h) + 1 \quad \text{and} \quad D_b = D(b) + 1 \quad (4.55)$$

where D_h and D_b are in the Kernighan method, and $D(h)$ and $D(b)$ are in our exchange method i.e.,

$$D_h = 2 \sum_{j=1}^n a_{hj}x_j - \sum_{j=1}^n a_{hj} + 1 \quad \text{and} \quad D_b = \sum_{j=1}^n a_{bj} - 2 \sum_{j=1}^n a_{bj}x_j + 1.$$

Proof. For $h \in H$,

$$\begin{aligned}
D_h &= \text{External cost} - \text{Internal cost} \\
&= \sum_{j \in B} c_{hj} - \sum_{j \in H} c_{hj} \\
&= \sum_{j \in B} a_{hj} - \left(\sum_{j \in H} a_{hj} - 1 \right) \quad \text{since } h \in H \text{ and } C = A - I \\
&= \sum_{j=1}^n a_{hj}x_j - \sum_{j=1}^n a_{hj}(1 - x_j) + 1 \\
&= 2 \sum_{j=1}^n a_{hj}x_j - \sum_{j=1}^n a_{hj} + 1.
\end{aligned}$$

For $b \in B$,

$$\begin{aligned}
 D_b &= \text{External cost} - \text{Internal cost} \\
 &= \sum_{j \in H} c_{bj} - \sum_{j \in B} c_{bj} \\
 &= \sum_{j \in H} a_{bj} - \left(\sum_{j \in B} a_{bj} - 1 \right) \quad \text{since } b \in B \text{ and } C = A - I \\
 &= \sum_{j=1}^n a_{bj}(1 - x_j) - \sum_{j=1}^n a_{bj}x_j + 1 \\
 &= \sum_{j=1}^n a_{bj} - 2 \sum_{j=1}^n a_{bj}x_j + 1.
 \end{aligned}$$

□

Theorem 27 *Under above conditions, the formulas (4.52) and (4.50) are equivalent, and the formula (4.54) is equivalent to the formula (4.51).*

Proof. Since $h_1 \in H$ and $b_1 \in B$ are exchanged, by (4.53) in lemma 25 and (4.55) in lemma 26,

$$\begin{aligned}
 g_{b_1 h_1} &= D(b) - D(h) + 2 - 2a_{b_1 h_1} \\
 &\Leftrightarrow g_{b_1 h_1} = D_{b_1} + D_{h_1} - 2a_{b_1 h_1} \\
 &\Leftrightarrow g_{b_1 h_1} = D_{b_1} + D_{h_1} - 2c_{b_1 h_1} \quad \text{since } b_1 \neq h_1.
 \end{aligned}$$

This shows that (4.52) is equivalent to (4.50). Now Let us show that new $D_h, D'_h = D_h + 2c_{hh_1} - 2c_{hb_1}$ and new $D_b, D'_b = D_b + 2c_{bb_1} - 2c_{bh_1}$ where $h \in H - \{h_1\}$ and $b \in B - \{b_1\}$ are equivalent to $D(i)' = D(i) + 2a_{ib_1} - 2a_{ih_1}$.

For $h \in H - \{h_1\}$ or $h \in H + \{b_1\} - \{h_1\}$, and $b \in B - \{b_1\}$ or $b \in B + \{h_1\} - \{b_1\}$,

$$a_{bh_1} = c_{bh_1}, \quad a_{bb_1} = c_{bb_1}, \quad a_{hh_1} = c_{hh_1}, \quad a_{hb_1} = c_{hb_1}.$$

Hence,

$$D'_b = D_b - 2c_{bh_1} + 2c_{bb_1}$$

$$\Leftrightarrow (D(b) + 1)' = D(b) + 1 - 2c_{bh_1} + 2c_{bb_1}$$

$$\Leftrightarrow D(b)' + 1 = D(b) + 1 - 2a_{bh_1} + 2a_{bb_1}$$

$$\Leftrightarrow D(b)' = D(b) - 2a_{bh_1} + 2a_{bb_1}.$$

Similarly, we have

$$D'_h = D_h - 2c_{hh_1} + 2c_{hb_1} \Leftrightarrow D(h)' = D(h) - 2a_{hh_1} + 2a_{hb_1}.$$

Hence, (4.54) \equiv (4.51). \square

Thus, the formula for gain in our exchange method is equivalent to the formula for gain in the Kernighan-Lin method.

6.4 Exchange Method Procedure

We will start with $H = \{i : x_i = 0\}$ and $B = \{i : x_i = 1\}$ and separate two stages.

First stage. we compute $D(j)$ for all j and find maximum of $D(b)$ for $b \in B$ and minimum of $D(h)$ for $h \in H$. Let $D(b^*) = \max\{D(b) : b \in B\}$ and $D(h^*) = \min\{D(h) : h \in H\}$.

Lemma 28 *If $D(b^*) - D(h^*) + 2 < 0$, then $g_{bh} < 0$ for all $(b, h) \in (B, H)$.*

Proof. Suppose $D(b^*) - D(h^*) + 2 < 0$. For any $b \in B$ and $h \in H$, since $D(b) \leq D(b^*)$ and $D(h) \geq D(h^*)$,

$$D(b) - D(h) + 2 \leq D(b^*) - D(h^*) + 2 < 0.$$

Since $a_{bh} = 0$ or $1 \forall (b, h) \in (B, H)$,

$$D(b) - D(h) + 2 < 0 \Rightarrow g_{bh} = D(b) - D(h) + 2 - 2a_{bh} < 0.$$

\square

Hence, if $D(b^*) - D(h^*) + 2 < 0$, then all gain is negative and so we have to stop.

Theorem 29 For $b \in B$, if $D(b) < D(h^*) - 2$ then for all $h \in H$, $g_{bh} < 0$. Also for $h \in H$, if $D(h) > D(b^*) + 2$ then for all $b \in B$, $g_{bh} < 0$.

Proof. Suppose $b \in B$ is given and $D(b) < D(h^*) - 2$. Then $D(b) - D(h^*) + 2 < 0$ and so for all $h \in H$,

$$g_{bh} = D(b) - D(h) + 2 - 2a_{bh} \leq D(b) - D(h^*) + 2 - 2a_{bh} < 0.$$

Similarly, for given $h \in H$, if $D(h) > D(b^*) + 2$ then for all $b \in B$, $g_{bh} < 0$. \square

Hence we consider only $\{b \in B : D(b) \geq D(h^*) - 2\}$ and $\{h \in H : D(h) \leq D(b^*) + 2\}$.

Theorem 30 If $D(b_*) < D(b^*) - 2$ for some $b_* \in B$ or $D(h_*) > D(h^*) + 2$ for some $h_* \in H$, then $g_{b_*h} < g_{b^*h^*}$ and $g_{bh_*} < g_{b^*h^*}$ for all $b \in B$ and $h \in H$.

Proof. Suppose $D(b_*) < D(b^*) - 2$, then for all $h \in H$,

$$\begin{aligned} g_{b_*h} &= D(b_*) - D(h) + 2 - 2a_{b_*h} \\ &< D(b^*) - 2 - D(h) + 2 - 2a_{b_*h} \\ &\leq D(b^*) - D(h) \\ &\leq D(b^*) - D(h) + 2 - a_{b^*h^*} \\ &= g_{b^*h^*}. \end{aligned}$$

Similarly, if $D(h_*) > D(h^*) + 2$, then for all $b \in B$, $g_{bh_*} < g_{b^*h^*}$. \square

Hence, we do not consider $b \in B$ such that $D(b) < D(b^*) - 2$ and $h \in H$ such that $D(h) > D(h^*) + 2$. That is, we consider only $\{b \in B : D(b) \geq D(b^*) - 2\}$ and $\{h \in H : D(h) \leq D(h^*) + 2\}$ to be exchanged. Therefore, by theorem 29 and theorem 30, we just consider

$$\{b \in B : D(b) \geq D(b^*) - 2, D(b) \geq D(h^*) - 2\}$$

and

$$\{h \in H : D(h) \leq D(h^*) + 2, D(h) \leq D(b^*) + 2\}.$$

Let us define for $i = 0, 1, 2$,

$$B_i = \{b \in B : D(b) = D(b^*) - i, D(b) \geq D(h^*) - 2\}$$

and

$$H_i = \{h \in H : D(h) = D(h^*) + i, D(h) \leq D(b^*) + 2\}.$$

Lemma 31 For $(b_i, h_j) \in (B_i, H_j)$, if $i + j > 2$, then $g_{b_0 h_0} > g_{b_i h_j}$, where $(b_0, h_0) \in (B_0, H_0)$.

Proof. Suppose $i + j > 2$ and $(b_i, h_j) \in (B_i, H_j)$. Then since $a_{b_0 h_0} = 0$ or $1 \forall i, j$,

$$\begin{aligned} g_{b_0 h_0} &= D(b_0) - D(h_0) + 2 - 2a_{b_0 h_0} \\ &= D(b_i) + i - (D(h_j) - j) + 2 - 2a_{b_0 h_0} \\ &\geq D(b_i) + i - D(h_j) + j \\ &> D(b_i) - D(h_j) + 2 \\ &\geq D(b_i) - D(h_j) + 2 - 2a_{b_i h_j} \\ &= g_{b_i h_j} \end{aligned}$$

□

If the maximum gain is positive, then select and exchange any pair of components which makes maximum gain and then go back to first stage with recalculated $D(i)$ by using (4.54). If the maximum gain is zero, then go to the second stage with the pairs of components whose gain is zero, say (b_{0_z}, h_{0_z}) , to make better selection for next iteration.

Second stage. In this stage, everything is temporary i.e. we will not bring any value in this stage to next iteration. We just check which pair of components of

zero gain can make positive gain in the next iteration. So, for efficiency of program, we will consider only $B_0 \cup B_1 \cup B_2$ and $H_0 \cup H_1 \cup H_2$ as B and H . In this stage, to choose which pair of components (b_{0_z}, h_{0_z}) whose gain is zero in a current iteration will be exchanged, we assume for each pair (b_{0_z}, h_{0_z}) , b_{0_z} and h_{0_z} are exchanged and then calculate the temporary $D(i)$, $D_T(i)$, for i in $H_T = \cup_{i=0}^2 H_i \cup \{b_{0_z}\} - \{h_{0_z}\}$ and $B_T = \cup_{i=0}^2 B_i \cup \{h_{0_z}\} - \{b_{0_z}\}$ by (4.54).

We are just looking for the largest possible positive gain. We call the positive gain in this stage possible positive gain, because it is not really positive gain in this iteration but the exchange of b_{0_z} and h_{0_z} will produce positive gain in the next iteration. Define

$$D_T(b^*) = \max\{D_T(b) : b \in B\} \text{ and } D_T(h^*) = \min\{D_T(h) : h \in H\}.$$

Then similarly to the first stage, if $D(b^*) < D_T(h^*) - 2$, then stop and so, for $i = 0, 1$, define

$$B_T = \{b \in B_T : D_T(b) = D_T(b^*) - i, D_T(b) \geq D_T(h^*) - 2\}$$

and

$$H_T = \{h \in H_T : D_T(h) = D_T(h^*) + i, D_T(h) \leq D_T(b^*) + 2\}.$$

For some (b_{0_z}, h_{0_z}) , if we find a possible positive new gain, then exchange the pair of components (b_{0_z}, h_{0_z}) and then go back to first stage. If (b_{0_z}, h_{0_z}) has no possible positive gain, then do it again with $(b_{0_{z+1}}, h_{0_{z+1}})$ until z is exhausted. If there is no positive gain for all of the pair of components, (b_{0_z}, h_{0_z}) , then stop the process.

But in practice, even if maximum gain is zero and for all (b_{0_z}, h_{0_z}) , there is no possible positive gain in a current iteration, we will continue the iterations by exchanging anyone of (b_{0_z}, h_{0_z}) 's. Because under the above circumstance, we may have some positive in a later iteration. Then we have to consider an infinite loop.

Lemma 32 *If there is a cycle, then for each iteration in the cycle, the maximum gain is zero.*

Proof. Suppose that $(\mathbf{1} - \mathbf{x}^{(p)})^T \mathbf{A} \mathbf{x}^{(p)} = (\mathbf{1} - \mathbf{x}^{(q)})^T \mathbf{A} \mathbf{x}^{(q)}$ with $q > p$. Since we do not take a negative gain, for each $j = p+1, \dots, q$,

$$(\mathbf{1} - \mathbf{x}^{(j-1)})^T \mathbf{A} \mathbf{x}^{(j-1)} \geq (\mathbf{1} - \mathbf{x}^{(j)})^T \mathbf{A} \mathbf{x}^{(j)}.$$

Hence,

$$(\mathbf{1} - \mathbf{x}^{(p)})^T \mathbf{A} \mathbf{x}^{(p)} \geq \dots \geq (\mathbf{1} - \mathbf{x}^{(p+1)})^T \mathbf{A} \mathbf{x}^{(p+1)} \geq \dots \geq (\mathbf{1} - \mathbf{x}^{(q)})^T \mathbf{A} \mathbf{x}^{(q)}.$$

Thus,

$$(\mathbf{1} - \mathbf{x}^{(j-1)})^T \mathbf{A} \mathbf{x}^{(j-1)} = (\mathbf{1} - \mathbf{x}^{(j)})^T \mathbf{A} \mathbf{x}^{(j)} \quad \forall j = p+1, p+2, \dots, q.$$

Therefore, every gain from $\mathbf{x}^{(p)}$ to $\mathbf{x}^{(q)}$ is zero. \square

Hence, to avoid infinite procedure, we need to stop if we have gain zero consecutively for a while.

Note in Figure 4.4 that $D(b^*)$ denotes maximum of $D(b)$ for $b \in B$, $D(h^*)$ denotes minimum of $D(h)$ for $h \in H$ and $g(b, h) = D(b) - D(h) + 2 - 2a_{bh}$ for all $b \in B$ and $h \in H$. $D_T(j)$ is the value of $D(j)$ for $j \in B^* = \cup_{i=0}^2 B_i$ and $H^* = \cup_{j=0}^2 H_j$ after changing two nodes. In practice, even if z is exhausted in the last step, we can go back to the beginning of this procedure with exchanging last pair of zero gain components (b_{0_z}, h_{0_z}) until it happen for a few times consecutively.

7 Stopping Criteria for a Local Minimizer after the Exchange Method

In the Exchange Method Procedure, if no components of the point is changed, then it is a local minimizer and so stop the program. If it is changed,, then we need to check that whether it is a local minimizer or not. So, we will go back to starting procedure.

Finally, we finished the whole processes to partition the given graph into two sets. Now we consider some improvement by exchanging sets instead of exchanging nodes, one from each partitioned graph.

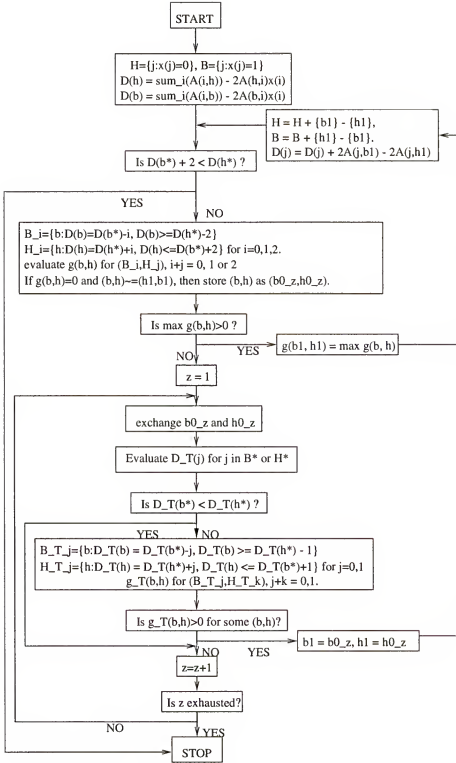


Figure 4.4: Exchange method for a local minimizer.

8 Exchange the Subsets from Partitioned Graph

We partitioned an initial graph into two subgraphs, V and W . In this phase, we now exchange some components of V and those of W to reduce the *edge-cut* between V and W . That is, we can reduce the *edge-cut* between V and W by swapping subsets V_1 of V and W_1 of W . Let

$$V = V_1 \cup V_2, \quad V_1 \cap V_2 = \emptyset, \quad W = W_1 \cup W_2 \text{ and } W_1 \cap W_2 = \emptyset$$

where $V_1 = \{i : x_{1i} = 1\}$, $W_1 = \{i : x_{2i} = 1\}$, $V \cup W = \{1, 2, \dots, n\}$ and $V \cap W = \emptyset$.

If the **edge-cut** between V_1 and V_2 , and between W_1 and W_2 is less than that between V_1 and W_2 , and between V_2 and W_1 , then the **edge-cut** between V and W will be reduced by exchanging V_1 and W_1 where $V_1 \subset V$, $W_1 \subset W$ and $|V_1| = |W_1|$. We use the gradient projection method with active set strategy as before. The implementation is almost the same as before.

CHAPTER 5 NUMERICAL RESULT

At first, we test our scheme on a small graph (see Figure 5.1) with the initial point $\mathbf{x}_0 = \mathbf{1}^T \frac{m}{n}$ to see the improvement of exchange method.

In Table 5.1 we will see the numerical result of Kernighan's method and exchange method (improved Kernighan's method) for the quadratic programming problem

$$\begin{aligned} \min f(\mathbf{x}), \quad & f(\mathbf{x}) = (\mathbf{1} - \mathbf{x})^T \mathbf{A} \mathbf{x} \\ \text{subject to } & \mathbf{0} \leq \mathbf{x} \leq \mathbf{1}, \mathbf{1}^T \mathbf{x} = m \end{aligned}$$

with random generated initial point, \mathbf{x}_0 with all 0 or 1 components.

In Table 5.1, note that $n(g = 0)$ denotes the maximum number of consecutive zero gain steps allowed, where $n(g = 0) = 0$ implies only positive gain is accepted, *i.e.*, Kernighan's method.

From Table 5.1, we can see that there is a big improvement between Kernighan's method and exchange method (improved Kernighan's method), but for the exchange method, there is not much difference between 2 consecutive zero gain steps and 4 consecutive zero gain steps.

In Table 5.2 with the initial point $\mathbf{x}_0 = \mathbf{1}^T \frac{m}{n}$, we compare the numerical result of

- (i) the gradient projection method, and
- (ii) the combination of (i) and exchange method for a local minimizer whose components are all 0 or 1 with up to 2 consecutive zero gain steps.

In Table 5.2 note that G.P.M. and E.M. denote respectively the gradient projection method and the exchange method (improved Kernighan's method).

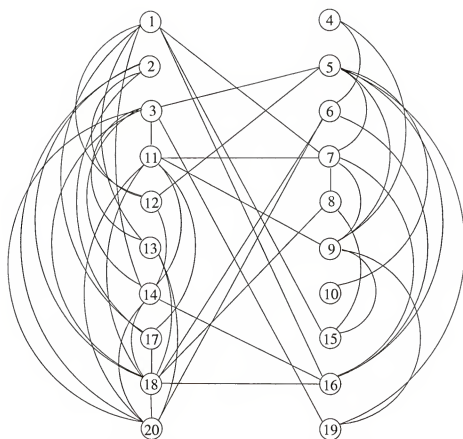


Figure 5.1: Example of graph with 20 vertices

Table 5.1: Comparison of Kernighan's method and the exchange method.

m	$f(x) \setminus n(g=0)$	0	2	4
2	3	885	1000	1000
	4	115		
3	5	874	1000	1000
	7	126		
4	7	592	850	850
	8	212		
	9	196	150	150
5	9	597	817	817
	11	325	183	183
	12	69		
	16	9		
6	10	586	666	671
	13	204	288	288
	14	116	5	
	15	39		
	16	44	41	41
	17	11		
7	11	561	629	629
	14	104	144	144
	15	212	227	227
	16	101		
	17	21		
	18	1		

m	$f(x) \setminus n(g=0)$	0	2	4
8	11	345	477	477
	13	70	70	70
	14	198	352	352
	15	93		
	16	189	69	69
	17	64	32	32
	18	31		
	19	10		
9	13	481	559	559
	14	308	407	423
	16	13		
	17	114	13	13
	18	61	21	5
10	19	13		
	13	212	804	804
	14	390		
	14	390		
	15	161	196	196
	17	77		
	18	134		
	19	26		

Table 5.2: Result of the combination of the gradient projection method, and the exchange method with the initial point $\mathbf{x} = \mathbf{1} \cdot \frac{m}{n}$.

m	global min.	G.P.M	G.P.M. & E.M.
2	3	3	3
3	5	5	5
4	7	7	7
5	9	9	9
6	10	10	10
7	11	12	11
8	11	13	11
9	13	19	14
10	13	14	13

In the Table 5.2, we can recognize that by using exchange method we have a lot better result. The significant improvement is for all m except $m = 9$, we got the global minimum.

Now, we test our scheme on reasonably big problems, linear programming and G-sets (<ftp://dollar.biz.uiowa.edu/pub/yyye/Gset/>) with the initial point which is the projection of the solution of the ball onto the convex set K . In the Tables 5.3 and 5.4, we put “★” for the better partition *i.e.*, for less *edge-cut* between two parts.

In Table 5.3, we can see the result of perfectly balanced bisection of our Quadratic Program and Metis for graph partitioning problems associated with linear programming. Note that the adjacency matrix is nonzero pattern for $\mathbf{A}\mathbf{A}^T$ where \mathbf{A} is the coefficient matrix for the linear program.

In Table 5.4, we have the result of perfectly balanced bisection of our Quadratic Program and Metis for G-set graphs.

Table 5.3: LP graph bisection test problems.

Problem	Vertices	Edges	Q P	Metis
Adlittle	56	328	88	88
Afiro	27	63	9	9
Beaconfd	173	2669	659 *	825
Vtp-base	198	1544	163	163
Scorpion	388	1713	22 *	23
Degen2	444	6855	1391 *	1445
ffff800	524	10091	1183 *	1249
Finnis	497	2771	142 *	217
Ganges	1309	7656	118	84 *
Perold	625	6303	423	423
Cycle	1903	27714	133	122 *

Table 5.4: G-Set graph bisection test problems.

Problem	Vertices	Edges	Q P	Metis
G1	800	19176	7653 *	7754
G3	800	19176	7609 *	7746
G4	800	19176	7647 *	7691
G5	800	19176	7632 *	7682
G14	800	4694	1129 *	1146
G15	800	4661	1140 *	1159
G16	800	4672	1121 *	1126
G17	800	4667	1089 *	1139
G51	1000	5909	1395 *	1427
G52	1000	5916	1406 *	1422
G53	1000	5914	1431 *	1457
G54	1000	5916	1399 *	1419
G43	1000	9990	3387 *	3395
G44	1000	9990	3409 *	3443
G45	1000	9990	3418 *	3460
G46	1000	9990	3380 *	3461
G47	1000	9990	3373 *	3455
G35	2000	11778	2878	2845 *
G36	2000	11766	2919	2846 *
G37	2000	11785	2882	2879 *
G38	2000	11779	2899 *	2902
G22	2000	19990	6843 *	6867
G23	2000	19990	6806 *	6862
G24	2000	19990	6835 *	6837
G25	2000	19990	6823 *	6886
G26	2000	19990	6833 *	6887

CHAPTER 6 CONCLUSION

We have found that the choice of initial guess plays an important role for the convergence to a local minimizer. In comparing our graph partitioning algorithm to the Metis package, the quality of min-cuts of our algorithm is better than that of Metis. The results presented here are relevant for the partitioning into two sets with unit weight of edges and vertices. We are considering an extension to weighted problems.

REFERENCES

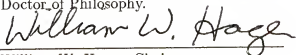
- [1] J.P.Aubin and A.Cellina. "Differential Inclusions." Springer-Verlag Berlin: 1984.
- [2] P. Berman, N. Koor, and P.M. Pardalos. "Algorithm for the Least Distance Problem." Piotr Berman and Nainan Koor, Computer Science Department, The Pennsylvania State University, University Park, PA 16804 USA. Panos M. Pardalos, Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611 USA. (1993).
- [3] D. P. Bertsekas. "Nonlinear Programming." Belmont, MA: Athena Scientific, (1995).
- [4] P. Brucker. "An $\mathcal{O}(n)$ algorithm for the quadratic knapsack problem." Operations Research Letters 3(1984), 163-166.
- [5] L. Hagen and A. Kahng. "Fast spectral methods for ratio cut and clustering." In Proceedings of IEEE International Conference on Computer Aided Design (1991), 10-13.
- [6] L. Hagen and A. Kahng. "A new approach to effective circuit clustering." In Proceedings of IEEE International Conference on Computer Aided Design (1992), 422-427.
- [7] W. W. Hager and Y. Krylyuk. "Graph Partitioning and Continuous Quadratic Programming." Department of Mathematics, University of Florida, Gainesville, FL 32611, March 7, 1998.
- [8] W. W. Hager. "Iterative Methods for Nearly Singular Linear Systems" Department of Mathematics, University of Florida, Gainesville, FL 32611, October 14, 1998.
- [9] W. W. Hager. "Applied Numerical Linear Algebra." Department of Mathematics, University of Florida, Gainesville, FL 32611, 1988.
- [10] R. Helgason, J. Kennington, and H. Lall. "A polynomial bounded algorithm for a single constrained quadratic program." Mathematical Programming 18(1980), 338-343.
- [11] G. Karypis and V. Kumar. "A Parallel Algorithm for Multilevel Graph Partitioning and Sparse Matrix Ordering." University of Minnesota, Department of Computer Science/Army HPC Research Center, Minneapolis, MN 55455, March 27, 1998.
- [12] G. Karypis and V. Kumar. "A fast and highly quality multilevel scheme for partitioning irregular graphs." University of Minnesota, Department of Computer Science, Minneapolis, MN 55455, March 27, 1998

- [13] B.W. Kernighan and S. Lin. "An efficient heuristic procedure for partitioning graphs." *Bell System Technical Journal* 49(1970), 291-307.
- [14] N. Kover and P.M. Pardalos. "An algorithm for a singly constrained class of quadratic programs subject to upper and lower bounds." *Mathematical Programming* 46(1990), 321-328.
- [15] T. Lengauer. "Combinatorial Algorithms for Integrated Circuit Layout." Chichester: John Wiley, 1990.
- [16] W. Li, P. M. Pardalos, and C. G. Han. "Gauss-Seidel Method for Least-Distance Problems" *Journal of Optimization Theory and Applications*: Vol.75, No.3. (December 1992), 487-500.
- [17] D. G. Luenberger. "Linear and Nonlinear Programming." Reading, MA: Addison-Wesley, 1984.
- [18] D. G. Luenberger. "Optimization by Vector Space Methods." New York, NY: John Wiley, 1968.
- [19] J. J. More and D.C. Sorensen. "Computing a trust region step." *SIAM Journal of Scientific and Statistical Computing* Vol.4, No.3. (September 1983), 553-572.

BIOGRAPHICAL SKETCH

Soonchul Park was born in Chungdo, Kyungsangbookdo, Korea, on January 6, 1964. He received his B.S degree in mathematics from Yonsei University, Wonju, Korea, in February 1992. In August 1994, he became a graduate student in the department of mathematics, University of Florida, from which he received his Ph. D. in mathematics in 1999.

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.



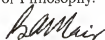
William W. Hager, Chairman
Professor of Mathematics

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.



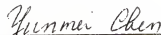
Gang Bao
Associate Professor of Mathematics

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.



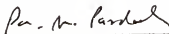
Bernard A. Mair
Professor of Mathematics

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.



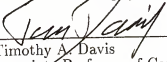
Yunmei Chen
Professor of Mathematics

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.



Panagote M. Pardalos
Professor of Industrial and Systems
Engineering

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.



Timothy A. Davis
Associate Professor of Computer and
Information Science and Engineering

This dissertation was submitted to the Graduate Faculty of the Department of Mathematics in the College of Liberal Arts and Sciences and to the Graduate School and was accepted as partial fulfillment of the requirements for the degree of Doctor of Philosophy.

August 1999

Dean, Graduate School